Journal of Machine Learning Research 26 (2025) 1-75 Submitted 9/24; Revised 4/25; Published 7/25

Early Alignment in Two-Layer Networks Training is a Two-Edged Sword

Etienne Boursier ETIENNE.BOURSIER@INRIA.FR
Université Paris-Saclay, CNRS, Inria, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France

Nicolas Flammarion NICOLAS.FLAMMARION@EPFL.CH
TML Lab, EPFL, Switzerland

Editor: Brian Kulis

Abstract

Training neural networks with first order optimisation methods is at the core of the empirical success of deep learning. The scale of initialisation is a crucial factor, as small initialisations are generally associated to a feature learning regime, for which gradient descent is implicitly biased towards *simple* solutions. This work provides a general and quantitative description of the early alignment phase, originally introduced by Maennel et al. (2018). For small initialisation and one hidden ReLU layer networks, the early stage of the training dynamics leads to an alignment of the neurons towards key directions. This alignment induces a sparse representation of the network, which is directly related to the implicit bias of gradient flow at convergence. This sparsity inducing alignment however comes at the expense of difficulties in minimising the training objective: we also provide a simple data example for which overparameterised networks fail to converge towards global minima and only converge to a spurious stationary point instead.

Keywords: Implicit Bias, Gradient Flow, ReLU Networks, Training Dynamics

1. Introduction

Artificial neural networks are nowadays used in numerous applications (He et al., 2016; Jumper et al., 2021). A part of their success originates from the ability of optimisation methods to find global minima, despite the non-convexity of the training losses; as well as the good generalisation performances obtained despite a large overparameterisation and an interpolation of the data (Zhang et al., 2021; Geiping et al., 2021; Liu et al., 2020). The understanding of these two generally admitted reasons yet remains very limited in the machine learning community. Recently, different lines of work advanced our comprehension of the empirical success of neural networks. First, Mei et al. (2018); Chizat and Bach (2018); Wojtowytsch (2020); Rotskoff and Vanden-Eijnden (2022) proved convergence of first order optimisation methods towards global minima of the training loss for idealised infinite width architectures. On the other hand, benign overfitting occurs in many different statistical models, i.e., the learnt estimator yields a small generalisation error despite interpolating the training data (Belkin et al., 2018; Bartlett et al., 2020; Frei et al., 2022; Tsigler and Bartlett, 2023).

This surprisingly good generalisation performance of neural networks is often attributed to the *implicit bias* of the used optimisation algorithms, that do select a specific global

minimum of the objective function (Neyshabur et al., 2014; Zhang et al., 2021). For linear neural networks, implicit bias has been thoroughly characterised (Ji and Telgarsky, 2019; Arora et al., 2019b; Yun et al., 2021; Min et al., 2021; Varre et al., 2023). In the presence of non-linear activations, e.g., ReLU, the implicit bias is much harder to characterise (Vardi and Shamir, 2021). In the classification setting, the learnt estimator is proportional to the min-norm margin classifier (Lyu and Li, 2019; Chizat and Bach, 2020). It is yet much more unclear in the regression case, for which it has only been characterised for specific data examples. In particular, Boursier et al. (2022) suggest that gradient flow is biased towards minimal norm interpolators; while other works suggest it induces sparsity in the representation of the network, in the sense it could be represented by a small number of neurons (Shevchenko et al., 2022; Safran et al., 2022; Chistikov et al., 2023). Although different (Chistikov et al., 2023), both notions of minimal norm and sparsity seem closely related (Parhi and Nowak, 2021; Stewart et al., 2023; Boursier and Flammarion, 2023). A similar implicit bias towards low rank solutions has been conjectured for matrix factorisation (Gunasekar et al., 2017; Arora et al., 2019a; Razin and Cohen, 2020).

Due to the non-convexity of the considered loss, the convergence point of training crucially depends on the choice of initialisation. Large initialisation is known to lead to the Neural Tangent Kernel (NTK) regime, for which gradient descent provably converges exponentially towards a global minimum of the training loss (Jacot et al., 2018; Du et al., 2018; Arora et al., 2019b). Unfortunately, this regime is also associated with lazy training, where the weights parameters only slightly change (Chizat et al., 2019). As a consequence, features are not learnt during training and can lead to a poor generalisation performance (Arora et al., 2019b).

On the other hand, smaller initialisations, such as in the mean field regime, yield a favorable implicit bias (Chizat and Bach, 2020; Jacot et al., 2021; Boursier et al., 2022), while still having some convergence guarantees towards global minima (Chizat and Bach, 2018; Wojtowytsch, 2020). However, this regime is more intricate to analyse and still lacks strong results on both convergence guarantees and implicit bias. In this objective, a recent part of the literature focuses on a complete description of the training dynamics, for both classification and regression, with specific data assumptions such as orthogonally separable (Phuong and Lampert, 2020; Wang and Pilanci, 2021), symmetric linearly separable (Lyu et al., 2021), orthogonal (Boursier et al., 2022), positively correlated (Wang and Ma, 2023; Chistikov et al., 2023; Min et al., 2024) or XOR-type data (Glasgow, 2024).

In particular, these works rely on a first early alignment phase, during which the neurons' weights all align towards a few key directions, while remaining small in norm and having no or little impact on the estimator prediction. In their seminal paper, Maennel et al. (2018) first described this key phenomenon, providing general heuristics for infinitesimal initialisation scale. This phase, specific to small initialisation and homogeneous activations, thus already induces some sparsity in the directions represented by the network and seems key to a final implicit bias with similar sparsity induced properties. This early alignment is not specific to the one hidden layer (Jacot et al., 2021) and has been empirically observed with more complex architectures and real data (Gur-Ari et al., 2018; Atanasov et al., 2021; Ranadive et al., 2023).

Contributions. Our contribution is twofold. First, we characterise the early alignment phenomenon for small initialisation, one hidden (leaky) ReLU layer networks trained with gradient flow in a general setup covering both classification and regression. A general study, i.e., holding for general datasets, of the so-called alignment has only been proposed by Maennel et al. (2018). This previous study however fails at precisely quantifying this alignment, since it follows heuristic arguments, for infinitely small initialisations. In opposition, we provide a finite time, macroscopic initialisation and rigorous analysis of the early alignment phase. As a consequence, our result (Theorem 1) can be directly applied to numerous previous works characterising the training dynamics of one hidden layer neural networks, to describe their first phase of dynamics.

Second, we apply this result to analyse the complete dynamics of training on a specific data example, for which gradient flow converges towards a spurious stationary point. In particular, our result implies that for small initialisation scales, interpolation might not happen at the end of training, even with an infinite number of neurons and infinite training time. This failure of convergence for largely overparameterised networks is suprising, as it goes in the opposite direction of previous works (Chizat and Bach, 2018; Wojtowytsch, 2020). This negative result highlights the importance of weights' omnidirectionality to reach global minima, a property that can be lost by early alignment for non-differentiable activations.

Overall, our work provides a general description of the early alignment phenomenon for small initialisations. This early phase presents clear benefits as it induces some sparsity of the network representation, that can be preserved along the whole trajectory. However, this benefit also comes at the expense of minimising the training loss, even on relatively simple datasets.

The concurrent works of Kumar and Haupt (2024); Tsoy and Konstantinov (2024) also provide a mathematical description of the early alignment, with the main differences that these results i) hold for a different set of assumptions (e.g. smooth activations or assuming a unique solution of the gradient flow); ii) do not provide a quantitative bound on the initialisation scale at which early alignment happens. As a consequence of the second point, their results do not hold in the limit of infinite width network. In contrast, our main Theorem 1 holds for this *mean field* limit, which is a key feature of our no-convergence result in Theorem 2.

2. Setting

We consider n data points $(x_k, y_k)_{k\in[n]}$ with features $x_k \in \mathbb{R}^d$ and labels $y_k \in \mathbb{R}$. We also denote by $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n\times d}$ the matrix whose rows are given by the input vectors. A two-layer neural network is parameterised by $\theta = (w_j, a_j)_{j\in[m]} \in \mathbb{R}^{m\times(d+1)}$, corresponding to the estimated function

$$h_\theta : x \mapsto \sum_{j=1}^{m} a_j \sigma(\langle w_j, x\rangle), \tag{1}$$

where σ is the (leaky) ReLU activation defined as $\sigma(x) \coloneqq \max(x, \gamma x)$ with $\gamma \in [0, 1]$. The parameters w_j and a_j respectively account for the hidden and output layer of the network.

Note that Equation (1) does not account for any bias term, as a simple reparameterisation of the features $\tilde{x} = (x, 1)$ allows to do so.

Training aims at minimising the empirical loss over the training dataset defined, for some loss function $\ell : \mathbb{R}^2 \to \mathbb{R}_+$, by

$$L(\theta) \coloneqq \frac{1}{n} \sum_{k=1}^{n} \ell(h_\theta(x_k), y_k).$$

As the limiting dynamics of gradient descent with infinitesimal learning rate, we study a solution of the following differential inclusion, for almost any $t \in \mathbb{R}_+$,

$$\frac{\mathrm{d}\theta^t}{\mathrm{d}t} \in -\partial_\theta L(\theta^t), \tag{2}$$

where $\partial_\theta L(\theta)$ is the Clarke subdifferential of L at θ (Clarke, 1990). Although the loss is not differentiable, the chain rule can still be applied for ReLU networks (Bolte and Pauwels, 2020, 2021), which is crucial to our analysis.

In the following, we consider a general loss function with minimal properties, covering both classical choices of square loss, $\ell(\hat{y}, y) = (\hat{y} - y)^2$, and logistic loss with binary labels, $\ell(\hat{y}, y) = \ln(1 + e^{-\hat{y}y})$.

Assumption 1. *For any $y \in \mathbb{R}$, the function $\hat{y} \mapsto \ell(\hat{y}, y)$ is differentiable. Moreover its derivative, denoted by $\partial_1 \ell(\cdot, y)$, is 1-Lipschitz and verifies $\partial_1 \ell(0, y) \neq 0$ for any $y \neq 0$.*

The existence of a global solution to Equation (2) is guaranteed (see, e.g., Aubin and Cellina, 2012, Chapter 2). However, such solutions can be non-unique since the loss function is not continuously differentiable in θ. In the following, *any* global solution of Equation (2) is considered.

Initialisation. The choice of initialisation is crucial when training two-layer neural networks, since the considered optimisation problem is non-convex. The m neurons of the neural network are here initialised as

$$(a_j^0, w_j^0) = \frac{\lambda}{\sqrt{m}} (\tilde{a}_j, \tilde{w}_j), \tag{3}$$

where $\lambda > 0$ is the scale of initialisation (independent of m) and $(\tilde{a}_j, \tilde{w}_j)$ are drawn i.i.d. such that almost surely[1]

$$|\tilde{a}_j| \geq \|\tilde{w}_j\| \text{ for any } j \in \mathbb{N}^*, \tag{4}$$

$$\text{and } \frac{1}{k} \sum_{j=1}^{k} \tilde{a}_j^2 \leq 1 \text{ for any } k \in \mathbb{N}^*. \tag{5}$$

The $\frac{1}{\sqrt{m}}$ factor in Equation (3) characterises the feature learning (or mean field) regime. In absence of this $\frac{1}{\sqrt{m}}$ term, no relevant feature would be learnt as it corresponds to the lazy regime (Chizat et al., 2019). Since the scale of a_j^0 can be controlled through λ, Equation (5) is compatible with any classical initialisation. In particular, it holds almost surely when $\tilde{a}_j$

1. Equation (5) is stated for any k (not just $k = m$), since we aim at stating results with constants that do not depend on the width m.

is bounded and it holds with high probability if $\tilde{a}_j$ is sub-Gaussian. As an example, any initialisation where

$$a_j^0 \in \left\{-\frac{\lambda}{\sqrt{m}}, \frac{\lambda}{\sqrt{m}}\right\} \quad \text{and} \quad w_j^0 \sim \mathcal{U}\left(B(0, \frac{\lambda}{\sqrt{m}})\right),$$

satisfies the above description.

2.1 Notations

We note $f(\lambda, t) = \mathcal{O}\left(g(\lambda, t)\right)$ if there exists a constant $c > 0$ that **only depends on the dataset**[2] $(x_k, y_k)_{k\in[n]}$ such that $|f(\lambda, t)| \leq c g(\lambda, t)$ on the considered set for λ and t. Occasionally, we note $f(\lambda, t) = \mathcal{O}_\alpha(g(\lambda, t))$ if the constant $c > 0$ depends on the dataset and an extra parameter α. Conversely, we note $f = \Omega(g)$ if $g = \mathcal{O}(f)$. We also note $f = \Theta(g)$ if both $f = \mathcal{O}(g)$ and $f = \Omega(g)$.

Along the paper, the detailed proofs are postponed to the appendix, for sake of readability.

3. Weight Alignment Phenomenon

This section aims at precisely quantifying the early alignment phenomenon in the setting of Section 2. For each individual neuron, Equation (2) rewrites

$$\frac{\mathrm{d}w_j^t}{\mathrm{d}t} \in a_j^t \mathfrak{D}_j^t \quad \text{and} \quad \frac{\mathrm{d}a_j^t}{\mathrm{d}t} \in \langle w_j^t, \mathfrak{D}_j^t \rangle, \tag{6}$$

$$\text{where } \mathfrak{D}_j^t = \mathfrak{D}(w_j^t, \theta^t) \coloneqq \Big\{ -\frac{1}{n}\sum_{k=1}^n \eta_k \partial_1 \ell(h_{\theta^t}(x_k), y_k) x_k \;\Big|\; \forall k \in [n], \eta_k \begin{cases} = 1 \text{ if } \langle w_j^t, x_k\rangle > 0 \\ = \gamma \text{ if } \langle w_j^t, x_k\rangle < 0 \\ \in [\gamma, 1] \text{ otherwise} \end{cases} \Big\}.$$

In the following, we also note by $D(w, \theta)$ the minimal norm subgradient, which is uniquely defined and happens to be of particular interest:

$$D(w, \theta) = \operatorname*{argmin}_{D \in \mathfrak{D}(w,\theta)} \|D\|_2 \,.$$

Note that the set $\mathfrak{D}_j^t$ only depends on the parameters through the estimated function h_{θ^t} and the activations of the neuron $A(w_j^t)$ where A is defined by

$$A: \begin{array}{l} \mathbb{R}^d \to \{-1, 0, 1\}^n \\ w \mapsto \left(\operatorname{sign}(\langle w, x_k\rangle)\right)_{k\in[n]} \end{array},$$

where $\operatorname{sign}(0) \coloneqq 0$ by convention. We thus also note in the following for any $u \in \{-1, 0, 1\}^n$: $\mathfrak{D}_u \coloneqq \mathfrak{D}(w, \mathbf{0})$ for any $w \in A^{-1}(u)$, since its definition does not depend on the choice of w. If $u \notin A(\mathbb{R}^d)$, we note $\mathfrak{D}_u = \emptyset$ by convention.

Our whole analysis relies on a first well known result, corresponding to the balancedness property (Arora et al., 2019b; Boursier et al., 2022).

2. It is yet independent of the number of neurons m.

Lemma 1 (Balancedness)**.** *For any $j \in [m]$ and $t \geq 0$, $(a_j^t)^2 - \|w_j^t\|^2 = (a_j^0)^2 - \|w_j^0\|^2$.*

Continuity of the neuron weights and Equation (4) then ensure that the sign of a_j^t remains constant during the whole training. We also make the following assumption on the data.

Assumption 2. *The data points (x_k, y_k) for any $k \in [n]$ are generated independently, following a distribution that is absolutely continuous with respect to the Lebesgue distribution on $\mathbb{R}^{d+1}$.*

Assumption 2 allows to avoid degenerate situations due to data. In particular, it is required to ensure the following Lemma 2, as well as Lemmas 9, 10 and 11 in the appendix.

Lemma 2. *If Assumptions 1 and 2 hold, then almost surely, for any $u \in \{-1, 0, 1\}^n$*

$$\mathfrak{D}_u \cap \left(\partial \overline{A^{-1}(u)} \cup -\partial \overline{A^{-1}(u)}\right) = \emptyset \text{ or } \mathbf{0} \in \mathfrak{D}_u,$$

*where $\overline{A^{-1}(u)}$ denotes the closure of $A^{-1}(u)$ and $\partial \overline{A^{-1}(u)}$ is the boundary **of the manifold** $\overline{A^{-1}(u)}$. Also, any family $(\partial_1 \ell(0, y_k) x_k)_k$ with at most d vectors is linearly independent.*

The boundary of the manifold $\partial \overline{A^{-1}(u)}$, which is different from the topological boundary (see, e.g., Tu, 2011, Section 22), is here given by

$$\partial \overline{A^{-1}(u)} = \Big\{ w \in \mathbb{R}^d \mid \forall k \in [n], \operatorname{sign}(\langle w, x_k \rangle) \begin{cases} \geq 0 \text{ if } u_k = 1 \\ = 0 \text{ if } u_k = 0 \\ \leq 0 \text{ if } u_k = -1 \end{cases} \text{ and } A(w) \neq u \Big\}.$$

Lemma 2 proves useful in our analysis, as it ensures that all the neuron dynamics at the end of the early alignment phase occur in the interior of the manifolds $A^{-1}(u)$, enabling to control these neurons. Following Maennel et al. (2018), Definition 1 introduces extremal vectors, which are key to the early alignment.

Definition 1. *For any $u \in \{-1, 0, 1\}^n$, the vector $D \in \mathfrak{D}_u$ is said **extremal** if $D \neq \mathbf{0}$ and $D \in -A^{-1}(u) \cup A^{-1}(u)$.*

Extremal vectors actually correspond to the critical points (up to rescaling) of the following function on the unit sphere

$$G : \begin{array}{l} \mathbb{S}_d \to \mathbb{R} \\ w \mapsto \langle w, D(w, \mathbf{0}) \rangle \end{array}. \tag{7}$$

The function G is piecewise linear. It is indeed linear on each activation cone $A^{-1}(u) \subset \mathbb{R}^d$. As a consequence, it has at most one critical point per cone. In general, the number of extremal vectors is even much smaller than the number of activation cones, since some of the cones do not include any critical point. Understanding the function G is crucial, since it is at the core of the early alignment phase.

Lemma 3. *If Assumptions 1 and 2 hold, there exists almost surely at least one extremal vector.*

In the early alignment phase, the parameters norm remains small so that $h_{\theta^t} \approx 0$. Meanwhile, the vectors $\frac{w_j^t}{a_j^t}$ approximately follow an ascending sub-gradient flow of G on the

unit ball of $\mathbb{R}^d$ (see Equation (8) in the proof sketch of Theorem 1). Since this directional movement happens much faster than the norm growth of the parameters, the vectors w_j^t end up being aligned in direction to the critical points of G, i.e., the extremal vectors. Theorem 1 precisely quantifies this phenomenon for every neuron satisfying Condition 1 below.

Condition 1. *The neuron $j \in [m]$ satisfies Condition 1 for $\alpha_0 > 0$ if both*

1. $\langle D(w_j^0, \mathbf{0}), \frac{w_j^0}{a_j^0} \rangle > -\sqrt{1-\alpha_0^2}\|D(w_j^0, \mathbf{0})\|$;
2. *for any $t \in \mathbb{R}_+$:* $\quad w_j^t = \mathbf{0} \implies w_j^{t'} = \mathbf{0}$ *for all $t' \geq t$.*

The meaning and necessity of this individual neuron condition is discussed further in Remarks 1 and 2 below.

Theorem 1. *If Assumptions 1 and 2 hold and the function G defined in Equation (7) does not admit a saddle point, then the following holds for any constant $\varepsilon \in (0, \frac{1}{3})$, $\alpha_0 > 0$ and initialisation scale $\lambda < \lambda^*_{\alpha_0}$ where $\lambda^*_{\alpha_0} > 0$ only depends[3] on the data $(x_k, y_k)_k$, α_0 and the activation parameter γ; with $D_{\max} \coloneqq \max_{w \in \mathbb{R}^d} \|D(w, \mathbf{0})\|$ and $\tau \coloneqq -\frac{\varepsilon \ln(\lambda)}{D_{\max}}$,*

(i) output weights do not grow large until τ:

$$\forall t \leq \tau, \forall j \in [m], |a_j^0|\lambda^{2\varepsilon} \leq |a_j^t| \leq |a_j^0|\lambda^{-2\varepsilon} \quad \text{and} \quad \|w_j^t\| \leq |a_j^t|.$$

(ii) Moreover, for any neuron j satisfying Condition 1 for α_0, $D(w_j^\tau, \mathbf{0})$ is either an extremal vector or $\mathbf{0}$, along which w_j^τ is aligned:

$$\|D(w_j^\tau, \mathbf{0})\| \geq \langle D(w_j^\tau, \mathbf{0}), \frac{w_j^\tau}{a_j^\tau} \rangle \geq \|D(w_j^\tau, \mathbf{0})\| - \mathcal{O}_{\alpha_0}\left(\lambda^{\frac{\|D(w_j^\tau, \mathbf{0})\|}{D_{\max}}\varepsilon}\right),$$

$$\text{or} \quad \frac{w_j^\tau}{\|w_j^\tau\|} = -\frac{D(w_j^\tau, \mathbf{0})}{\|D(w_j^\tau, \mathbf{0})\|}.$$

Also, the direction towards which w_j^τ is aligned corresponds to a local maximum (resp. minimum) of G if $a_j^0 > 0$ (resp. $a_j^0 < 0$).

Theorem 1 describes for a small enough scale of initialisation the early alignment phase, which happens during a time of order $\varepsilon \ln(\frac{1}{\lambda})$ at the beginning of the training dynamics. First, the neurons all remain of small norm during this phase—while the term $\lambda^{-2\varepsilon}$ grows large as $\lambda \to 0$, the other term $|a_j^0|$ also scales in λ, making their product bounded and arbitrarily small as $\lambda \to 0$.

Second, neurons end up aligned towards a few key directions, given by extremal vectors. There are indeed few such directions: as a first observation, there is at most one extremal vector $D(w, \mathbf{0})$ per activation cone, and there are at most $\mathcal{O}\left(\min(3^n, n^d)\right)$ such cones (see e.g., Cover, 1965, Theorem 4). In general, the number of extremal vectors is even much smaller. For example, studies describing the complete parameters dynamics (Phuong and Lampert, 2020; Lyu et al., 2021; Boursier et al., 2022; Chistikov et al., 2023; Min et al., 2024; Wang and Ma, 2023) all count either one or two extremal vectors—with

3. The exact value of $\lambda^*_{\alpha_0}$ is given by Equation (21) in Appendix D.

the exception of Glasgow (2024), where the population loss counts 4 extremal vectors (we refer to Appendix G for more details about this fact). The subsequent work of Boursier and Flammarion (2024) even goes beyond, by studying the extremal vectors of G when the number of data points grows to infinity. In particular, they showed that for some linear data model, there are only two extremal vectors for large number of training samples. A general understanding of the number of extremal vectors yet remains open. We believe this quantization of represented directions to be closely related to the implicit bias of first order optimisation methods and elaborate further on this aspect in Section 6.

Note that some neurons w_j^τ are not aligned towards extremal vectors but instead have $D(w_j^\tau, \mathbf{0}) = \mathbf{0}$. Thanks to the first point of Lemma 11 in Appendix D, this means that the neuron is deactivated with all data points x_k. As a consequence, it does not move anymore during training and has no impact at all on the estimated function h_{θ^t}, i.e., these neurons can be ignored after the early alignment phase.

The complete proof of Theorem 1 is given in Appendix D and is sketched below, at the end of this section.

Remark 1. *The first point of Condition 1 only bounds away from -1 the alignment of w_j^t with $D(w_j^t, \mathbf{0})$. When $\alpha_0 \to 0$, this covers all the neurons with probability 1. However, when fixing $\alpha_0 > 0$ and let m go to infinity, a (small) fraction of neurons does not satisfy this condition. These neurons are hard to control, as their alignment speed is arbitrarily slow at the beginning of the procedure (see Equation (8)): they can then take an arbitrarily large time before being aligned to some extremal vector.*

Remark 2. *Neurons such that $w_i(t) = \mathbf{0}$ at some time can spontaneously leave $\mathbf{0}$ in a way that cannot be controlled—both in the time at which it leaves, and in the direction at which it does so—due to the multiplicity of subgradient flow solutions of Equation (2). The second point in Condition 1 then restricts the analysis to natural solutions, by assuming that as soon as a neuron is $\mathbf{0}$, it does not move anymore. This is what generally happens in practice for ReLU activations, where we fix $\sigma'(0) = 0$ in common implementations. Another way to ensure this point is to consider a balanced initialisation $|a_j^0| = \|w_j^0\|$. In that case, a simple Grönwall argument allows to show that both a_j^t and w_j^t never cancel, automatically guaranteeing the second point of Condition 1. Adding weight decay to the optimisation scheme (with any choice of regularisation parameter) would also ensure the second point of Condition 1.*

A significant assumption in Theorem 1 is that G has no saddle point. While G can have saddle points in complex data cases, we stress that the absence of saddles for G is also a significant possibility and covers all previous works describing the complete parameters dynamics (Phuong and Lampert, 2020; Lyu et al., 2021; Boursier et al., 2022; Chistikov et al., 2023; Min et al., 2024; Wang and Ma, 2023; Glasgow, 2024), as well as any 2-dimensional data.[4] As a consequence, the first phase of training in these works is fully grasped by Theorem 1. The additional technical contribution of Glasgow (2024) for the first phase lies in further bounding the difference between gradient flow and SGD during this early alignment phase and we detail further in Appendix G how our results can be applied to the XOR data setting studied by Glasgow (2024). We provide in Appendix E

4. We refer to Lemma 15 for a proof of this fact.

an adapted version of Theorem 1 that holds in the presence of saddle points, but requires a stronger condition on the neurons.

The presence of saddle points is much harder to handle in general, as neurons can evolve arbitrarily slowly near saddle points. Providing a finite time convergence for such neurons then does not seem possible. Additionally, non trivial phenomena can happen in the presence of saddle points that are due to the non-continuity of the loss gradient. Notably some non-zero neurons can spontaneously leave their activation regions in a way that cannot be uniformly controlled, due to the multiplicity of gradient flow solutions of Equation (2). In Appendix E, we again consider specific gradient flow solutions to control such phenomena, via Condition 2.

Recall that Assumption 2 is only required so that Lemmas 2, 9, 10 and 11 hold. A deterministic version of both Theorems 1 and 3 is then possible without Assumption 2, if we ensure that Lemmas 2, 9, 10 and 11 all simultaneously hold.

The concurrent works of Tsoy and Konstantinov (2024); Kumar and Haupt (2024) provide a complementary characterisation of this early alignment phase. Notably Tsoy and Konstantinov (2024) characterize, for smooth activations, the early alignment as well as the first growth of neurons cluster occurring after this alignment. As an artifact of their analysis, they also require an odd data dimension d. On the other hand, Kumar and Haupt (2024) describe the early alignment for more general network architectures, assuming a unique gradient flow solution.

In contrast to our work, their results do not quantify the initialisation scale $\lambda^*_{\alpha_0}$ at which this early alignment phase happens. Theorem 1 indeed provides a quantitative bound on such a scale, given by Equation (21) in Appendix D. Importantly, this bound does not depend on the network width m. This point is crucial in Section 4 as it leads to results that hold for any value of m and thus remain valid in the mean field limit when $m \to \infty$. Our Theorem 2 in Section 4 being valid in the mean field limit is a crucial point, since it implies that the seminal result by Chizat and Bach (2018) of global convergence of overparameterised networks does not extend to the non-smooth case. This relation to Chizat and Bach (2018) is discussed further in Section 6.

Although our scale threshold $\lambda^*_{\alpha_0}$ is independent of m, it still depends in the data. It is unclear how this threshold scales with quantities of interest, such as the number of training samples n or the data dimension d. Given the generality of the data considered here (Assumption 2), we cannot expect a more precise description of this dependence. However, we believe such a description to be both of interest and feasible for typical data models. In particular, the subsequent work of Boursier and Flammarion (2024) studies the evolution of this scale threshold with both n and d for a linear data model with Gaussian inputs.

Sketch of proof. The first point of the proof follows from a simple Grönwall inequality argument on Equation (6), using the fact that $|a_j^t| \geq \|w_j^t\|$ for any t by balancedness.

From the first point, $h_{\theta^t}(x_k) = \mathcal{O}\left(\lambda^{2-4\varepsilon}\right)$ during the early alignment phase. As a consequence, we can derive the following approximate ODE almost everywhere:

$$\frac{\mathrm{d}\langle \mathbf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} = \|D(w_j^t, \mathbf{0})\|^2 - \langle \mathbf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 - \mathcal{O}\left(\lambda^{2-4\varepsilon}\right), \tag{8}$$

where $\mathsf{w}_j^t = \frac{w_j^t}{a_j^t}$ is the direction of the neuron j. If $a_j^t > 0$ (resp. $a_j^t < 0$), this equality corresponds to an approximate projected gradient ascent (resp. descent) of G on the unit ball.

In a first time, thanks to Condition 1, $|\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle|$ is bounded away from $\|D(w_j^t, \mathbf{0})\|$, which ensures that $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ is increasing at a rate $\Omega_{\alpha_0}(1)$, until either $D(w_j^{t_2}, \mathbf{0}) = \mathbf{0}$ or $|\langle \mathsf{w}_j^{t_2}, D(w_j^{t_2}, \mathbf{0})\rangle|$ is close to $\|D(w_j^t, \mathbf{0})\|$ for some time $t_2 = \mathcal{O}_{\alpha_0}(1)$.

The former case implies that the neuron is deactivated for the remaining of the training, and thus $D(w_j^\tau, \mathbf{0}) = \mathbf{0}$. The second case implies w_j^t is close in direction to $D(w_j^t, \mathbf{0})$. Moreover since $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ is increasing at the start of training and G has no saddle points, $D(w_j^t, \mathbf{0})$ corresponds to a local maximal direction of G if $a_j^0 > 0$ (and minimal if $a_j^0 < 0$).

As a consequence, we can show by studying the local maxima (or minima) of G that if $\mathsf{w}_j^{t_2}$ is negatively correlated with $D(w_j^{t_2}, \mathbf{0})$, it is positively proportional to $-D(w_j^{t_2}, \mathbf{0})$. From there, w_j^t stays aligned with $-D(w_j^{t_2}, \mathbf{0})$ until τ and only changes in norm. If instead $\mathsf{w}_j^{t_2}$ is positively correlated with $D(w_j^{t_2}, \mathbf{0})$, then we have a stability result (Lemma 13 in Appendix D.5) showing that

$$A(w_j^t) = A(w_j^{t_2}) \quad \text{for any } t \in [t_2, \tau]. \tag{9}$$

In other words, $D(w_j^t, \mathbf{0})$ is constant on $[t_2, \tau]$, so that Equation (8) is close to an ODE of the form $f'(t) = c - f(t)^2$ on this interval. It then implies, by Grönwall comparison, exponential convergence of $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ towards $\|D(w_j^t, \mathbf{0})\|$ and allows to conclude the second point of Theorem 1. □

Although these arguments are rather intuitive, their rigorous proof is tedious. Indeed, the estimated function h_{θ^t} is not exactly 0. It is only close to 0 during the early alignment phase, so that the neurons' dynamics are not exactly controlled by the vectors $D(w_j^t, \mathbf{0})$, but small perturbations of these vectors. As a consequence, we have to control carefully all the perturbed dynamics. In particular, showing the stability of the critical manifolds, i.e., Equation (9), is quite technical and is done in Appendices D.4 and D.5. The fact that stability is still preserved when slightly perturbing the vectors $D(w_j^t, \mathbf{0})$ largely relies on Assumption 2.

4. Convergence Towards Spurious Stationary Points

As explained in Section 6 below, the early alignment phenomenon has an interesting impact, especially on the implicit bias of gradient descent. However, it can also lead to undesirable, counter-intuitive results. This section aims to demonstrate that, in simple data examples, the early alignment phenomenon can be the cause of largely overparameterised neural networks converging to spurious stationary points. Such a result is somewhat surprising, as most of the current literature suggests that with a large enough number of neurons in the mean field regime, the learnt estimator converges towards a zero training loss function. This discrepancy with previous results is discussed further in Section 6.

We consider in this section the particular case of regression with ReLU activation: $\sigma(x) = \max(0, x)$ and

$$\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2. \tag{10}$$

We consider the following 3 points data example ($n = 3$ in this section).

Assumption 3. *The data is given by* 3 *points* $(x_k, y_k) \in \mathbb{R}^3$*, for some* $\eta > 0$*,*

$$x_1 \in (-1, -1 + \eta] \times [1, 1 + \eta] \text{ and } y_1 \in [1, 1 + \eta];$$
$$x_2 \in [-\eta, \eta] \times [1 - \eta, 1 + \eta] \text{ and } y_2 \in (0, \eta];$$
$$x_3 \in [1 - \eta, 1) \times [1, 1 + \eta] \text{ and } y_3 \in [1, 1 + \eta].$$

This assumption corresponds to a non-zero measure set of $\mathbb{R}^{3\times 3}$, so that it cannot be considered as a specific degenerate case. Also setting all the second coordinates of the x_k to 1 is possible and would correspond to the univariate dataset shown in Figure 1 below, with bias terms in the hidden layer of the network.

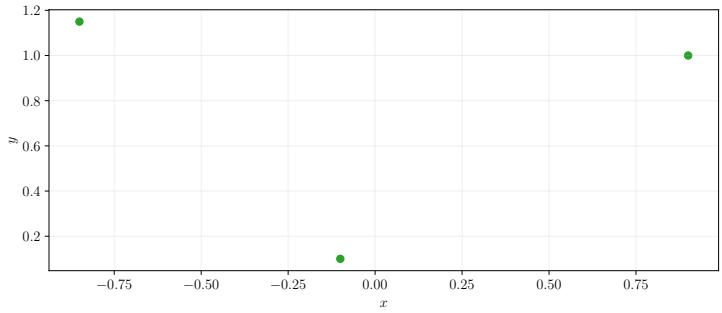


Figure 1: example of univariate data verifying Assumption 3 with $\eta = \frac{1}{6}$. The second coordinate of the features x_k is here fixed to 1 to take into account bias terms of the neural network.

Theorem 2. *Let* $\beta^* := \left(X^\top X\right)^{-1} X^\top y$ *be the ordinary least squares estimator of the data. If Assumption 3 holds with* $\eta < \frac{1}{6}$*, then there exists some constant* $\tilde{\lambda} = \Theta(1)$ *such that for any* $\lambda < \tilde{\lambda}$ *and* $m \in \mathbb{N}$*, the parameters* θ^t *converge to some* θ_∞ *such that*

$$h_{\theta_\infty}(x_k) = x_k^\top \beta^* \text{ for any } k \in [n].$$

In particular, it satisfies $\lim_{t\to\infty} L(\theta^t) > 0$.

A more refined version of Theorem 2, stated in Appendix F, even states that the learnt estimator is very close the positive part of the linear, ordinary least squares estimator, when taking $\lambda \to 0$. With the data example given by Assumption 3, the linear estimator corresponding to β^* does not fit all the data, while a simple two-neurons network can. In conclusion, although the data still seems easy to fit, Theorem 2 implies that for an initialisation scale smaller than $\tilde{\lambda}$, the learnt parameters converge towards a spurious local stationary point.

The scale of initialisation λ does not depend on the width m, but solely on the training data. In particular, Theorem 2 even holds in the limit $m \to \infty$. Even though this result might seem to contradict known global convergence results in the infinite width limit at first hand (Chizat and Bach, 2018; Wojtowytsch, 2020), it is actually compatible with these results and highlights the importance of some assumptions that appear to be technical in nature (differential activation or infinite data) to get these global convergence results. A further discussion on this aspect is given in Section 6.

The proof of Theorem 2 describes the complete dynamics of the parameters θ during training. This dynamics happens in three distinct phases, described in detail in Section 4.1. In particular, the first phase of the proof is a direct consequence of our early alignment result given by Theorem 1. Because of this alignment, the positive neurons then nearly behave as a single neuron, which is positively correlated with all data points. From there, adding a neuron with negative output weights only increases the training loss, for any direction of the weight. As a consequence, it becomes impossible to activate a new direction while training and the network remains equivalent to a single neuron one until the end of training.

Remark 3. *The condition $\eta < \frac{1}{6}$ does not seem to be tight and is only needed for analytical reasons here. In Section 5, a similar result is empirically observed for a larger η. We provide in Appendix A more general data assumptions for which Theorem 2 still holds.*

4.1 Sketch of Proof

This section provides a sketch of proof for Theorem 2. In particular, the proof relies on the description of a three phases training dynamics as follows.

1. The first phase corresponds to the early alignment one, after which all the neurons with positive output weights are aligned with the single extremal vector.

2. During the second phase, the neurons with positive output weights all grow in norm, while staying aligned. After this phase, the estimated function h_{θ^t} is already very close to the single neuron network represented by β^*.

3. During the last phase, a local Polyak-Łojasiewicz (PL) argument allows to state that the parameters θ do not significantly move and converge at an exponential speed towards the linear estimator β^*.

The first and second phase are a mere consequence of the fact that for our data example, there is a unique extremal vector given by $\frac{1}{n}\sum_{k=1}^{n} y_k x_k$. The third phase on the other hand is a consequence of both that i) all positive neurons are aligned and behave as a single neuron at the end of the early alignment phase; ii) the data is designed so that adding any negative ReLU neuron to the linear estimator will only increase the training loss. In consequence, no additional neuron will be "created" during this third phase and the estimator will remain equivalent to β^*.

4.1.1 Phase 1

The first phase of the training dynamics is the early alignment one and can be fully described by Theorem 1. First define the sets

$$\mathcal{I} \coloneqq \left\{ i \in [m] \mid a_i^0 > 0 \text{ and } \exists k \in [n], \langle w_i^0, x_k \rangle > 0 \right\},$$
$$\mathcal{N} \coloneqq \left\{ i \in [m] \mid a_i^0 < 0 \right\}.$$

Also define the vectors

$$D^t = -\frac{1}{n} \sum_{k=1}^{n} (h_{\theta^t}(x_k) - y_k) x_k \quad \text{and} \quad D^* = \frac{1}{n} \sum_{k=1}^{n} y_k x_k.$$

An important property of the considered dataset is that there is a single extremal vector, given by D^*, and all data points are pairwise positively correlated. During the first alignment phase, all the neurons with positive output weights then end up aligned with D^* as stated by Lemma 4 below.

Lemma 4 (Phase 1)**.** *Under Assumption 3 with $\eta < \frac{1}{6}$, for any $\varepsilon \in (0, \frac{1}{3})$, $\lambda < \tilde{\lambda}$ and $\tau = \frac{-\varepsilon \ln(\lambda)}{D_{\max}}$.*

1. *Positive neurons are aligned with D^*: $\forall i \in \mathcal{I}, \langle D^*, \frac{w_i^\tau}{a_i^\tau} \rangle = \|D^*\| - \mathcal{O}(\lambda^\varepsilon)$ and $a_i^0 \leq a_i^\tau \leq a_i^0 \lambda^{-2\varepsilon}$.*
2. *Negative neurons' norm decreased: $\forall i \in \mathcal{N}, a_i^0 \leq a_i^\tau \leq 0$.*
3. *Remaining neurons do not move during the whole training: $\forall i \notin \mathcal{I} \cup \mathcal{N}, \forall t \in \mathbb{R}_+, w_i^t = w_i^0$ and $a_i^t = a_i^0$.*

Lemma 4 is a direct application of Theorem 1, up to additional minor remarks, that are specific to Assumption 3. It does not allow to control the directions of the neurons in $\mathcal{N}$. Such a control is actually not needed since the norm of the neurons in $\mathcal{N}$ stays close to 0 in the following phases.

4.1.2 Phase 2

At the end of the first phase, the neurons in $\mathcal{I}$ are aligned with D^* and positively correlated with all the points x_k. From there, their norm grows during the second phase, while the norm of neurons in $\mathcal{N}$ remains close to 0. While the neurons in $\mathcal{I}$ grow in norm, their direction also changes. Controlling both their direction and norm simultaneously is intricate and split into two subphases (2a and 2b) detailed below. At the end of this norm growth and change of direction during the second phase, the group of positive neurons almost behave as the optimal linear regressor β^*. In other words with τ_3 the end of the second phase, we have

$$\sum_{i \in \mathcal{I}} a_i^{\tau_3} w_i^{\tau_3} \approx \beta^* \quad \text{and} \quad \forall i \in \mathcal{I}, \forall k \in [n], \langle w_i^{\tau_3}, x_k \rangle > 0.$$

The second inequality implies that each neuron $i \in \mathcal{I}$ behaves linearly with the data points despite the ReLU activation.

Phase 2a. The subphase 2a is the first part of the phase 2. During this part, the norms of the neurons in $\mathcal{I}$ grow until reaching a small, but $\Theta(1)$ threshold given by ε_2. We choose ε_2 small enough, so that during this phase $D^t = D^* + \mathcal{O}(\varepsilon_2)$. As a consequence, the dynamics of the neurons during this part is not much different than in the early alignment phase. This behavior leads to Lemma 5 below.

Lemma 5 (Phase 2a)**.** *Consider ε, τ fixed by Lemma 4. Under Assumption 3 with $\eta < \frac{1}{6}$, there exist constants $c', \varepsilon_2^* = \Theta(1)$ such that for any $\varepsilon_2 \in [c'\lambda^{\frac{\varepsilon}{2}}, \varepsilon_2^*]$ and $\lambda < \tilde{\lambda}$, there is some $\tau_2 \in (\tau, \infty)$ such that*

1. *positive neurons norm reach the threshold: $\sum_{i\in\mathcal{I}}(a_i^{\tau_2})^2 = \varepsilon_2$;*
2. *positive neurons are positively correlated with all data points: $\forall i \in \mathcal{I}, k \in [n], \langle w_i^{\tau_2}, x_k\rangle = \Omega(1)$;*
3. *positive neurons are nearly aligned: $\forall i, j \in \mathcal{I}, \langle \mathsf{w}_i^{\tau_2}, \mathsf{w}_j^{\tau_2}\rangle = 1 - \mathcal{O}\left(\lambda^{1-\varepsilon}\right)$;*
4. *negative neurons' norm decreased: $\forall i \in \mathcal{N}, a_i^0 \leq a_i^{\tau_2} \leq 0$.*

Note that the third point of Lemma 5 only states that the positive neurons are pairwise aligned, while Lemma 4 stated that they were aligned with D^*. This is because at the time τ_2, the positive neurons might not be aligned anymore with D^* at a precision level λ. The direction of the positive neurons might indeed have slightly moved away from D^*, but all the positive neurons kept a common direction during this phase. This property proves useful in the following, as we want the neurons in $\mathcal{I}$ to behave as a single neuron, i.e., we want them all pairwise aligned.

Phase 2b. At the end of the first part, given by the Phase 2a, the positive neurons' have a norm ε_2 and are on the verge of exploding in norm. The Phase 2b corresponds to this explosion until the positive neurons almost correspond to the optimal linear regressor β^*. More precisely for some $\varepsilon_3 > 0$, the Phase 2b ends when

$$\left\| \beta^* - \sum_{i\in\mathcal{I}} a_i^{\tau_3} w_i^{\tau_3} \right\| \leq \varepsilon_3.$$

Lemma 6 below states that this condition is reached at some point, while the negative neurons remain small in norm and the positive neurons are still positively correlated with all data points.

Lemma 6 (Phase 2b)**.** *Consider $\varepsilon, \varepsilon_2, \tau_2$ fixed by Lemma 5. Under Assumption 3 with $\eta < \frac{1}{6}$, there exist constants $c_3 = \Theta(1)$ and $\tilde{\varepsilon}_2^* = \Theta(1)$ depending only on the data, such that if $\varepsilon_2 \leq \tilde{\varepsilon}_2^*$ and $\varepsilon_3 \geq \lambda^{c_3\varepsilon\varepsilon_2}$ and $\lambda < \tilde{\lambda}$, after some time $\tau_3 \in [\tau_2, +\infty)$,*

1. *positive neurons behave as the optimal linear regressor: $\left\|\beta^* - \sum_{i\in\mathcal{I}} a_i^{\tau_3} w_i^{\tau_3}\right\| \leq \varepsilon_3$;*
2. *positive neurons are still nearly aligned: $\forall i, j \in \mathcal{I}, \langle \mathsf{w}_i^{\tau_2}, \mathsf{w}_j^{\tau_2}\rangle = 1 - \mathcal{O}(\lambda^{\varepsilon})$;*
3. *positive neurons are positively correlated with all data points: $\forall i \in \mathcal{I}, \forall k \in [n]$, $\langle \frac{w_i^{\tau_3}}{a_i^{\tau_3}}, x_k\rangle = \Omega(1)$;*

4. negative neurons' norm is small: $\forall i \in \mathcal{N}, a_i^0 \lambda^{-\varepsilon} < a_i^{\tau_3} \leq 0$.

The main argument in the proof of Lemma 6 is that the Phase 2b happens in a time of order[5] $\mathcal{O}\left(\frac{-\ln(\varepsilon_3)}{\varepsilon_2}\right)$. Since we choose λ very small with respect to $\varepsilon_2, \varepsilon_3$, this time can be seen as very short. In comparison, the phase 2a ends after a time of order $-\ln(\lambda)$. In this short amount of time, the norm of the negative neurons cannot grow significantly enough and the positive neurons cannot significantly disalign. Thanks to the latter and the specific structure of the data, the positive neurons remain positively correlated with all data points during the whole second phase.

4.1.3 Phase 3

The third phase is only theoretical since it describes infinitesimal movement towards the convergence point β^*. For this phase, we define for $\varepsilon_4, \delta_4 > 0$,

$$\tau_4 \coloneqq \inf \left\{ t \geq \tau_3 \;\middle|\; \|\theta^t - \theta^{\tau_3}\|_2 \geq \varepsilon_4 \text{ or } \exists i \in \mathcal{I}, k \in [n] \langle \mathsf{w}_i^t, x_k \rangle \leq \delta_4 \|x_k\| \right\}.$$

Lemma 7 (Phase 3)**.** *If Assumption 3 holds with $\eta < \frac{1}{6}$, consider $\varepsilon, \varepsilon_2, \varepsilon_3, \tau_3$ fixed by Lemma 6. There are positive constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^* = \Theta(1)$ such that if we also have $\lambda < \tilde{\lambda}$, $\varepsilon_3 < \varepsilon_3^*$, $\varepsilon_4 = \varepsilon_4^*$ and $\delta_4 \leq \delta_4^*$, then $\tau_4 = \infty$. Moreover, there then exists θ^∞ such that*

1. *the parameters converge towards some limit:* $\lim_{t \to \infty} \theta^t = \theta^\infty$;
2. *all the neurons of this limit behave linearly with the training data:* $\forall i \in [m], (\forall k \in [n], \langle w_i^\infty, x_k \rangle \geq 0)$ *or* $(\forall k \in [n], \langle w_i^\infty, x_k \rangle \leq 0)$;
3. *the active neurons correspond to the optimal linear estimator:* $\sum_{\substack{i \in [m] \\ \forall k \in [m], \langle w_i^\infty, x_k \rangle \geq 0}} a_i^\infty w_i^\infty = \beta^*$.

Similarly to Chatterjee (2022), the proof of Lemma 7 relies on a local PL condition. The condition is a bit trickier here, since it is with respect to the loss obtained by a spurious stationary point instead of a global minimum. Getting this local PL condition requires to also control the negative neurons. Controlling them is intricate but possible by showing that they either act as a linear operator or that they decrease in norm.

Once we derive the local PL condition (see Lemma 22 in Appendix F.4), the remaining of the proof encloses the dynamics of the parameters in a small compact set and then shows exponential convergence towards some parameter θ^∞ satisfying the conditions of Lemma 7, which implies that

$$\forall k \in [n], h_{\theta^\infty}(x_k) = \langle \beta^*, x_k \rangle,$$
$$\forall x \in \mathbb{R}^d, h_{\theta^\infty}(x) = \langle \beta^*, x \rangle_+ + \mathcal{O}\left(\varepsilon_3 + \varepsilon_4\right).$$

5. Experiments

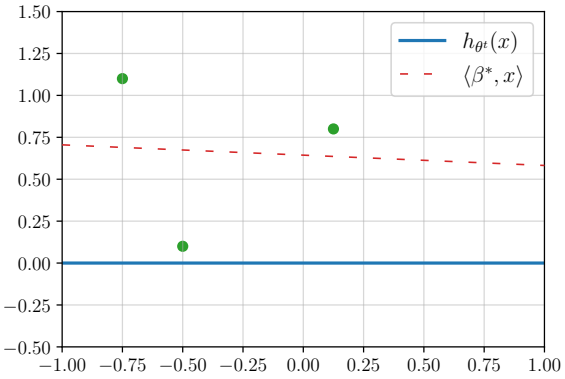
This section empirically illustrates the results of Theorems 1 and 2. The considered dataset does not exactly fit the conditions of Theorem 2 to illustrate that Assumption 3 with $\eta < \frac{1}{6}$

5. The vector $\sum_{i \in \mathcal{I}} a_i^t w_i^t$ is indeed approximating the gradient flow of the linear regression of the data, with a learning rate lower bounded by ε_2.

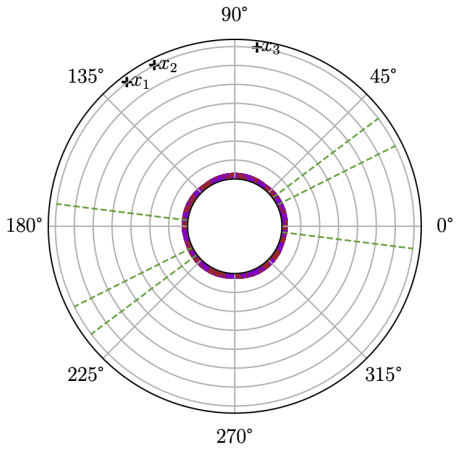
is only needed for analytical purposes. The dataset is however similar to datasets satisfying Assumption 3 (see e.g., Figure 1) in the sense that all three data points are positively correlated, with positive labels; and the middle point is below the optimal linear regressor. The data points are here represented as unidimensional, since their second coordinate is fixed to 1 to take into account bias terms in the hidden layer of the neural network.

For the given set of data points, we train a one-hidden ReLU network with gradient descent over the mean square loss given by Equation (10). Experimental details and additional experiments are provided in Appendix B. The code and animated versions of the figures are also available at `github.com/eboursier/early_alignment`.

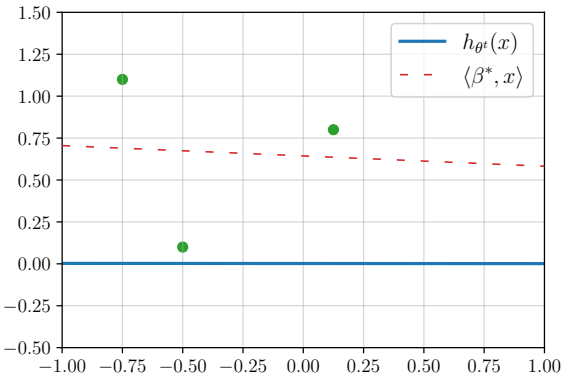
Figure 2 illustrates the training dynamics over time. The left column represents the estimated function $h_{\theta^t}(x)$ at different training times; while the right column represents the 2-dimensional repartition of the network weights w_j^t in polar coordinates. In the latter, the inner circle corresponds to 0 norm and is shifted away from the origin to accurately observe the early alignment phenomenon. Each star corresponds to a single neuron $w_j^t \in$



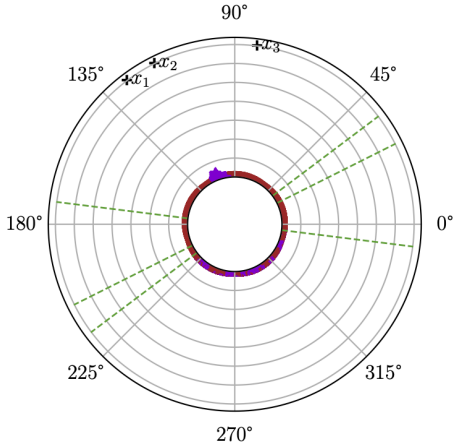
(a) Estimated function at initialisation (iteration 0)



(b) Weights' repartition at initialisation

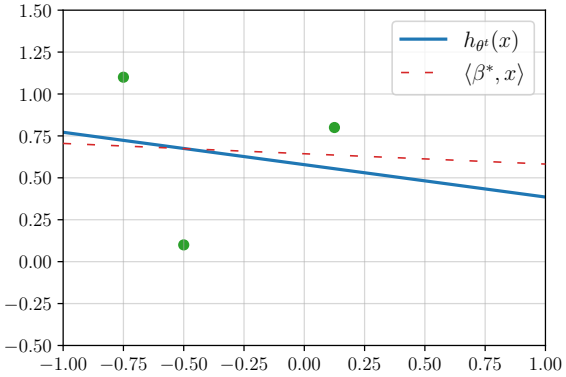


(c) Estimated function after early alignment (iteration 3000)

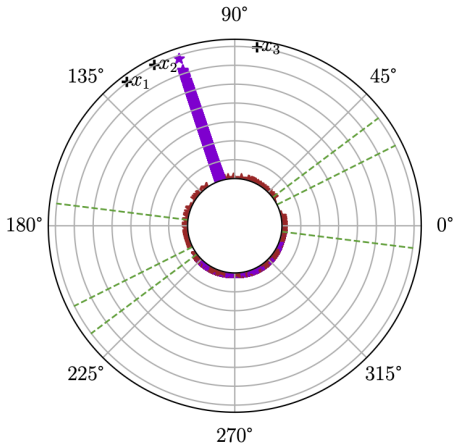


(d) Weights' repartition after early alignment

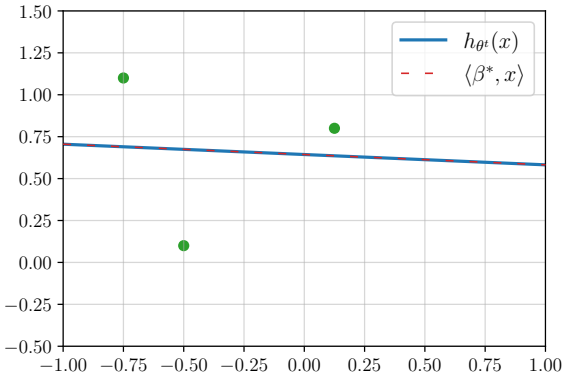
Figure 2: Training dynamics (part 1/2).



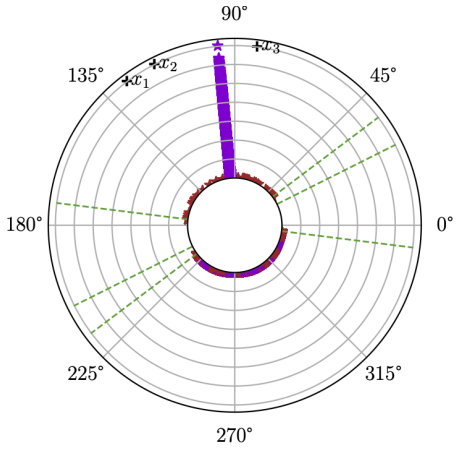
(e) Estimated function during neurons' growth (iteration 6500)



(f) Weights' repartition during neurons' growth



(g) Estimated function at convergence (iteration 2×10^6)



(h) Weights' repartition at convergence

Figure 2: Training dynamics (part 2/2).

$\mathbb{R}^2$, represented in purple (resp. red) if its associated output weight a_j^t is positive (resp. negative). The data points directions' x_k are represented by black crosses and the dotted green lines delimit the different activation cones $A^{-1}(u)$. We recall that the weights w_j^t are two-dimensional, as their second coordinate corresponds to bias terms.

Figures 2a and 2b show the parameters at initialisation. Given the small initialisation scale, the neurons are all nearly 0 norm and the estimated function is close to the zero function. Moreover, the neurons' directions are uniformly spread over the whole space.

Figures 2c and 2d then show the state of the network after the early alignment phase. In particular, all the positive neurons (in purple) are either aligned with the single extremal vector $\frac{1}{n}\sum_{k=1}^{n} y_k x_k$ or deactivated with all data points. The latter then never move from their initial position. Meanwhile, the neurons' norm is still close to 0 so that the estimated function is still nearly 0. Actually, Figures 2c and 2d precisely show the end of Phase 2a (see Lemma 5), where the positive neurons are on the verge of exploding in norm. Indeed, we

can see on Figure 2d a small purple bump: the positive neurons just start to considerably grow in norm.

Figures 2e and 2f show the learnt parameters during the norm growth of the positive neurons. They can be seen as the end of the Phase 2, where the learnt estimator is close enough to the optimal linear one. While the positive neurons considerably grew in norm, the negative neurons remained small in norm and are actually irrelevant to the learnt estimator.

Figures 2g and 2h finally illustrate the learnt parameters at convergence. The learnt estimator converged to the optimal linear one, as predicted by Theorem 2. In Figure 2h, all the neurons (including the negative ones) are located in only two activation cones: either they are activated with all the data points, or they are activated with no data point. This confirms our theoretical analysis and in particular the description given in the proof of Lemma 21 in Appendix D. Another interesting observation is that the final direction of the positive neurons changed from their direction at the end of the early alignment (see difference between Figures 2d and 2h). As explained in Section 4.1.2, this is one of the challenge in analysing the complete dynamics of the parameters: simultaneously controlling the norm growth and change of direction of the neurons is technically intricate.

6. Discussion

Early alignment and implicit bias. Theorem 1 states that neurons satisfying Condition 1 end up aligned towards a few key directions, given by extremal vectors, at the end of the early alignment phase. We believe this quantization of represented directions to be at the origin of the implicit bias of (stochastic) gradient descent. Many recent works indeed suggest that gradient descent is implicitly biased towards small rank hidden weights matrix W or, equivalently, a sparse number of directions represented by the neurons (Shevchenko et al., 2022; Safran et al., 2022). Other works evidence that the implicit bias can often be characterised by minimal norm interpolator (Lyu and Li, 2019; Chizat and Bach, 2020), which happen to be closely related to this notion of sparse represented directions (Parhi and Nowak, 2021; Stewart et al., 2023; Boursier and Flammarion, 2023). The early alignment phase seems to confirm such a behavior, since it enforces alignment of weights towards a small number of directions, even with an omnidirectional initialisation.

Another interesting point that can be made from both Theorem 2 and the experiments in Section 5 and Appendix B is that even if the training procedure does not manage to perfectly fit all the data, there still seems to be an implicit regularisation of the estimator at convergence. Specifically, for the 3 points example given in Theorem 2 and Section 5, the network does not manage to fit the data. It actually ends up being equivalent to a single neuron network, while two neurons are necessary to fit the data. However, there still seems to be some implicit regularisation at play, as the learnt estimation is very simple. Even more striking is that the learnt estimator represents the best possible way to fit the data with a single neuron, which is here given by the optimal linear estimator. A similar observation is made in Appendix B.2, where the network ultimately becomes equivalent to a 5 neurons network, but 8 neurons are required to fit the data. Moreover, it also seems that the obtained estimator is given by the best way to fit the data points with only 5 neurons here. Although these claims are pure intuitions and empirical observations, we believe that the stationary points reached in our different examples are only bad from an optimisation

point of view, but could still yield good generalisation behaviors. Such an intuition has been theoretically confirmed in a subsequent work for a linear data model (Boursier and Flammarion, 2024).

Omnidirectionality of the weights. Omnidirectionality is the property that the weights $(w_j^t, \text{sign}(a_j^t))$ cover all possible directions of $\mathbb{R}^d \times \{-1, 1\}$, in the limit of infinite width $m \to \infty$. This property has been used to show that two layer infinite width networks converge to a global minimum of the loss. In particular with an omnidirectional initialisation, omnidirectionality is preserved along the whole training if either the activation is differentiable (Chizat and Bach, 2018) or with ReLU and an infinite dataset (Wojtowytsch, 2020). This property allows to never get stuck in a bad local minimum, since there is always a direction along which increasing the neurons' norm decreases the training loss.

On the other hand, the spurious convergence result of Section 4 is based on the fact that the early alignment phase can cause loss of omnidirectionality of the weights.[6] At the end of the early alignment phase, the neurons with positive output weights only cover a very small cone around the sole extremal vector, as can be observed in Figure 2. This result does not contradict Chizat and Bach (2018); Wojtowytsch (2020), since the activation is non-differentiable and the data finite in our setting. It actually highlights how fragile and crucial is the omnidirectionality to guarantee convergence towards global minima in the mean field regime.

Also, Theorem 1 does not necessarily imply a loss of omnidirectionality. The weights can still cover all possible directions after the early alignment phase. Theorem 1 instead only implies that a large fraction of them is concentrated around a few key directions. Even with omnidirectional weights, the key directions thus have a much larger number of neurons with respect to other directions. We believe this discrepancy between the directions to be sufficient for getting an implicit bias towards small rank weights matrix.

Scale of initialisation. A strong feature of Theorem 2 is that the scale initialisation $\tilde{\lambda}$ does not depend on the network width m. The only dependence on m is given by the $\frac{1}{\sqrt{m}}$ scaling. The absence of this scaling corresponds to the lazy regime, where convergence towards global minimum happens, yet without feature learning, which can yield bad generalisation on unseen data (Chizat et al., 2019).

Obtaining Theorem 2 with a threshold $\tilde{\lambda}$ that depends on the width m is technically easier. However, such a setting would not be compatible with the infinite width network regimes considered by Chizat and Bach (2018); Wojtowytsch (2020). On the other hand when fixing λ and taking $m \to \infty$, our result lies in the infinite width setting. Moreover, our initialisation regime can be made compatible with the result of Wojtowytsch (2020) by choosing $|a_j^0| = \|w_j^0\|$ for all j, as explained at the bottom of their second page. As explained above, the only reason that our result does not contradict Chizat and Bach (2018); Wojtowytsch (2020) is that we considered non-differentiable activation with finite data.

A conclusion of Theorems 1 and 2 is that choosing a small initialisation scale should yield a better implicit bias towards low rank hidden weights matrix, but at the risk of

6. We stress in the paragraph below that our choice of initialisation is compatible with Chizat and Bach (2018); Wojtowytsch (2020).

converging towards spurious stationary points (still with low rank matrix). On the other hand, choosing a large initialisation scale can affect generalisation on new, unseen data (Chizat et al., 2019). In practice, an intermediate scale might thus be the best trade-off to yield both convergence to global minima and small generalisation errors. Yet on the theoretical side, analysing such intermediate regimes is even harder than extremal ones.

Nature of the stationary point. Until now, we only described the convergence point reached in the no convergence example of Theorem 2 as a spurious stationary point. Its exact nature, i.e., whether it is a saddle point or a spurious local minimum actually depends on the choice of initialisation. If the initialisation counts *perfectly balanced* neurons with $|a_j^0| = \|w_j^0\|$, the convergence point θ^∞ of the parameters, described by Theorem 4 in the Appendix, can include zero neurons ($a_j^\infty = 0$, $w_j^\infty = \mathbf{0}$), which corresponds to saddle points of the loss. We can indeed slightly perturb such a neuron in a good direction, so that it decreases the loss.

If on the other hand we only consider *dominated neurons* with $|a_j^0| > \|w_j^0\|$, the convergence point θ^∞ does not include zero neurons, as a_j^∞ is non-zero for any j. From there, computations similar to the proof of Lemma 22, along with the description of θ^∞ by Theorem 4, yield that the convergence point θ^∞ is actually a local minimum of the loss.

Generality of Theorem 2. Theorem 2 states that for some cases of data, convergence towards spurious stationary points happens. First note that Assumption 3 describes a non-zero measure set for any $\eta > 0$, so that this data example is not fully degenerate. Moreover, Theorem 2 is proven for $\eta < \frac{1}{6}$, but similar observations are empirically made for datasets that do not satisfy Assumption 3 (see Section 5).

More generally, we believe that this case of bad convergence happens on many different data examples. First, Assumption 3 is needed for three technical conditions, detailed in Appendix A. One of them allows to show that the whole network behaves as a single neuron after the early alignment. Yet on more complex data examples, the network can still get stuck after having created several neurons, as can be observed in Appendix B. Having a precise description of the dynamics then becomes more intricate, but the reason for the failure of training is the same: the loss of weights omnidirectionality hinders the growth of neurons in a good direction. We believe that such *bad convergence* could happen in many low-dimensional settings (in the sense $n \gtrsim d$). Indeed in such cases, the early alignment should lead to a sparse number of represented directions at its end, given by the number of critical points of the alignment function G. We conjecture this number of directions to be independent of n in the low-dimensional setting for typical structured data models (such as the ones that can be found in Bach, 2017, Table 1). Such an independence would then lead to an impossibility of fitting all data points, since the network would become equivalent to a smaller network with less than n neurons.

Besides its scale, the initialisation has the property of balancedness given by Equation (4) here. This property allows to fix the sign of the output weights a_j^t. Without it, some of the weights a_j^t change their sign during the early alignment phase. Because of that, the omnidirectionality of the weights $(w_j, \text{sign}(a_j))$ on $\mathbb{R}^d \times \{-1, 1\}$ is still preserved when considering unbalanced initialisations in the example of Theorem 2. The parameters θ^t thus converge towards a global minimum for unbalanced initialisation with the data of Assumption 3.

We yet believe that this omnidirectionality is not generally preserved with unbalanced initialisation, as the differentiability of the loss $L(\theta)$ is the general argument for the conservation of the omnidirectionality (Chizat and Bach, 2018; Wojtowytsch, 2020) and would still not be satisfied. This intuition is empirically confirmed in Appendix B.2, where omnidirectionality of the weights is not preserved and the network fails at finding a global minimum, despite an unbalanced initialisation. Theoretically studying examples where unbalanced initialisations fail to find global minima however is more complex and left open for future work.

Theorem 2 focuses on a regression task. Stewart et al. (2023) suggest that finding global minima is easier for classification than regression tasks and could be a reason for the empirical success of binning (i.e., recasting a regression task as a classification one). Whether a similar failure of training is possible with classification tasks is left open for future work. We believe it is indeed possible, since the early alignment phase still happens and can lead to the loss of omnidirectionality, which is also key in classification settings.

Additionally, Theorem 2 is specific to the ReLU activations. This particular choice of activation is crucial to its proof, as it implies that *all* positive neurons either align in a short time towards the extremal vector (point 1 in Lemma 4), or are not relevant (point 2 in Lemma 4). Using leaky ReLU activations, neurons in $[m] \setminus (\mathcal{I} \cup \mathcal{N})$ become relevant; and some of them move at an arbitrarily slow rate during the first phase. As a consequence, omnidirectionality would be preserved during the early alignment phase and convergence towards an interpolator still happens.[7] We yet allege that failure of convergence can still happen with leaky ReLU activations, yet at the expense of a slightly more intricate data example where omnidirectionality can be lost.

7. Conclusion

This work provides a first general and rigorous quantification of the early alignment phenomenon, that happens for small intialisations in one hidden layer (leaky) ReLU networks and was first introduced by Maennel et al. (2018). We believe that such a result can be directly used in future works that study the training dynamics of gradient flow for specific data examples. In particular, we apply this result to describe the training trajectory on a 3 points example. Despite the simplicity of this example, gradient flow fails at learning a global minimum and only converges towards a spurious stationary point (corresponding to the least squares regressor), even with an infinite number of neurons. This example thus illustrates the duality of the early alignment phenomenon: although it induces some sparsity in the learnt representation, which is closely related to the implicit bias of gradient flow, it also comes with a risk of failure in minimising the training loss.

Acknowledgments

The authors thank Lenaïc Chizat for insightful discussions. This work was partially funded by an unrestricted gift from Google and by the Swiss National Science Foundation (grant number 212111).

7. Theorem 2 still holds for leaky ReLU activations if we allow $\tilde{\lambda}$ to depend on m.

References

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. *Advances in Neural Information Processing Systems*, 32, 2019a.

Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019b.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2021.

J-P Aubin and Arrigo Cellina. *Differential inclusions: set-valued maps and viability theory*, volume 264. Springer Science & Business Media, 2012.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel J Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. *Advances in neural information processing systems*, 31, 2018.

Jérôme Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. *Advances in Neural Information Processing Systems*, 33:10809–10819, 2020.

Jérôme Bolte and Edouard Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*, 188: 19–51, 2021.

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. *Advances in Neural Information Processing Systems*, 36:57795–57824, 2023.

Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer relu networks. *arXiv preprint arXiv:2410.02348*, 2024.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

Sourav Chatterjee. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022.

Dmitry Chistikov, Matthias Englert, and Ranko Lazic. Learning a neuron by a shallow relu network: Dynamics and implicit bias for correlated inputs. *Advances in Neural Information Processing Systems*, 36:23748–23760, 2023.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3): 326–334, 1965.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

Jonas Geiping, Micah Goldblum, Phil Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2021.

Margalit Glasgow. SGD finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the XOR problem. In *The Twelfth International Conference on Learning Representations*, 2024.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Shengchao Liu, Dimitris Papailiopoulos, and Dimitris Achlioptas. Bad global minima exist and sgd can reach them. *Advances in Neural Information Processing Systems*, 33:8543–8552, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2019.

Kaifeng Lyu, Zhiyuan Li, Runzhe Wang, and Sanjeev Arora. Gradient descent on two-layer nets: Margin maximization and simplicity bias. *Advances in Neural Information Processing Systems*, 34:12978–12991, 2021.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Hancheng Min, Salma Tarmoun, Rene Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7760–7768. PMLR, 18–24 Jul 2021.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer reLU networks with small initialization. In *The Twelfth International Conference on Learning Representations*, 2024.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

Mary Phuong and Christoph H Lampert. The inductive bias of relu networks on orthogonally separable data. In *International Conference on Learning Representations*, 2020.

Omkar Ranadive, Nikhil Thakurdesai, Ari S. Morcos, Matthew L Leavitt, and Stephane Deny. On the special role of class-selective neurons in early training. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

Itay Safran, Gal Vardi, and Jason D Lee. On the effective number of linear regions in shallow univariate relu networks: Convergence guarantees and implicit bias. *Advances in Neural Information Processing Systems*, 35:32667–32679, 2022.

Alexander Shevchenko, Vyacheslav Kungurtsev, and Marco Mondelli. Mean-field analysis of piecewise linear solutions for wide relu networks. *The Journal of Machine Learning Research*, 23(1):5660–5714, 2022.

Lawrence Stewart, Francis Bach, Quentin Berthet, and Jean-Philippe Vert. Regression as classification: Influence of task formulation on neural network features. In *International Conference on Artificial Intelligence and Statistics*, pages 11563–11582. PMLR, 2023.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24:123–1, 2023.

Nikita Tsoy and Nikola Konstantinov. Simplicity bias of two-layer networks beyond linearly separable data. In *International Conference on Machine Learning*, 2024.

Loring W Tu. Manifolds. In *An Introduction to Manifolds*, pages 47–83. Springer, 2011.

Gal Vardi and Ohad Shamir. Implicit regularization in relu networks with the square loss. In *Conference on Learning Theory*, pages 4224–4258. PMLR, 2021.

Aditya Vardhan Varre, Maria-Luiza Vladarean, Loucas Pillaud-Vivien, and Nicolas Flammarion. On the spectral bias of two-layer linear networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Mingze Wang and Chao Ma. Understanding multi-phase optimization dynamics and rich nonlinear behaviors of relu networks. *Advances in Neural Information Processing Systems*, 36:35654–35747, 2023.

Yifei Wang and Mert Pilanci. The convex geometry of backpropagation: Neural network gradient flows converge to extreme points of the dual convex program. In *International Conference on Learning Representations*, 2021.

Stephan Wojtowytsch. On the convergence of gradient descent training for two-layer relu-networks in the mean field regime. *arXiv preprint arXiv:2005.13530*, 2020.

Chulhee Yun, Shankar Krishnan, and Hossein Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix

Table of Contents

Appendix A. More General Assumptions for Section 4

In this section, we provide more general assumptions than Assumption 3 that still lead to the spurious convergence result of Theorem 2.

Assumption 4. *For all k, k', $\langle x_k, x_{k'}\rangle > 0$, $y_k > 0$. Also, $n \geq d$ and any family $(x_k)_k$ with at most d vectors is linearly independent.*

The goal of this assumption is to have a single extremal vector, so that the early alignment leads to a concentration of (almost) all neurons towards the same direction. We assume the matrix $X \in \mathbb{R}^{n\times d}$ to be full rank with $n \geq d$ for the sake of simplicity. Note that the neurons do not move in the orthogonal space of $\mathrm{Span}(x_1, \dots, x_n)$ with gradient methods, so that our results can be extended to the case of rank deficient X (or $n < d$).

Assumption 5. *For any $k \in \mathcal{K}$, $y_k\|x_k\|^2 > \sqrt{\sum_{k'=1}^n y_{k'}^2}\sqrt{\sum_{k'\neq k}^n \langle x'_k, x_k\rangle^2}$, where*

$$\mathcal{K} \coloneqq \left\{k \in [n] \mid \exists w \in \mathbb{R}^d \setminus \{\mathbf{0}\}, \langle w, x_k\rangle = 0 \text{ and } \forall k' \in [n], \langle w, x_{k'}\rangle \geq 0\right\}.$$

This assumption is needed to assume that in the dynamics of a single neuron network, this neuron stays activated with all the data points ($\langle w, x_k\rangle > 0$ for any k) along its trajectory. It is needed so that our estimator behaves as a linear regression in the whole second phase of training.

Define in the following for any $u \in \{-1, 0, 1\}^n$ and for β^* as in Theorem 2,

$$\tilde{D}_u^{\beta^*} \coloneqq \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{u_k=1}(y_k - \langle \beta^*, x_k\rangle)x_k, \tag{11}$$

$$\mathfrak{D}_u^{\beta^*} \coloneqq \left\{\frac{1}{n}\sum_{k=1}^n \eta_k(y_k - x_k^\top\beta^*)x_k \;\middle|\; \forall k \in [n], \eta_k \begin{cases} \in [0,1] \text{ if } \langle w_j^t, x_k\rangle = 0 \\ = 1 \text{ if } \langle w_j^t, x_k\rangle > 0 \\ = 0 \text{ otherwise}\end{cases}\right\}.$$

Assumption 6. *For any $k \in \mathcal{K}$, $\langle \beta^*, x_k\rangle < y_k$.*
Moreover, for any $u \in A(\mathbb{R}^d)$, such that $\exists k, k' \in [n], u_k = 1$ and $u_{k'} = -1$, there exists $\delta_u > 0$ such that both

1. $\inf_{D\in\mathfrak{D}_u^{\beta^*}} u_k\langle D, x_k\rangle > 0$ *for any $k \in \mathcal{K}$ such that both $u_k \neq 0$ and* $\inf_{w\in A^{-1}(u)} |\langle \frac{w}{\|w\|}, \frac{x_k}{\|x_k\|}\rangle| \leq \delta_u$;
2. *for any $\delta \in (0, \delta_u)$,* $\inf_{w\in A_\delta^{-1}(u)} \langle \tilde{D}_u^{\beta^*}, \frac{w}{\|w\|}\rangle > 0$, *where*

$$A_\delta^{-1}(u) \coloneqq \left\{w \in A^{-1}(u) \mid \exists k, k' \in \mathcal{K},\ s.t.\ \langle \frac{w}{\|w\|}, \frac{x_k}{\|x_k\|}\rangle \geq \delta \text{ and } \langle \frac{w}{\|w\|}, \frac{x_{k'}}{\|x_{k'}\|}\rangle \leq -\delta\right\}.$$

Although technical, this assumption is equivalent to assuming that the least squares regression estimator β^* only gets worse, when a ReLU neuron with a negative output weight is added to it. This ensures that after the second phase of training, where the estimator is arbitrarily close to the least squares, the neurons with negative output weights do not grow in norm.

In our three points example, the two conditions stated in Assumption 6 are equivalent and mean that for the two extreme points, $\langle \beta^*, x_k\rangle < y_k$.

Lemma 8. *If Assumption 3 holds with $\eta < \frac{1}{6}$, then the dataset $(x_i, y_i)_{i=1,2,3}$ satisfies Assumptions 4, 5 and 6.*

Proof. 1) First check that Assumption 4 holds. Simple computations indeed lead to $\langle x_k, x_{k'} \rangle > 0$ if $\eta < 1$. Moreover, we indeed have $y_k > 0$. Also, x_1, x_2, x_3 are pairwise linearly independent for $\eta < 1/4$.

2) Some cumbersome but direct computations (that we here skip for sake of readibility) show that if $\eta < \frac{1}{6}$, we can then write

$$\begin{aligned} x_2 &= \alpha_1 x_1 + \alpha_3 x_3 \\ &\text{with } \alpha_1 > \eta \text{ and } \alpha_3 > 0. \end{aligned} \tag{12}$$

From there, we can choose $w_1 \neq \mathbf{0}$ such that $\langle x_1, w_1 \rangle = 0$ and $\langle x_3, w_1 \rangle > 0$. Equation (12) then directly leads to $\langle x_2, w_1 \rangle > 0$. By definition, this yields $1 \in \mathcal{K}$. Using similar arguments, $3 \in \mathcal{K}$.

Assume for some $w \neq \mathbf{0}$, $\langle w, x_2 \rangle = 0$ with $\langle w, x_1 \rangle \geq 0$ and $\langle w, x_3 \rangle \geq 0$. Equation (12) then implies that all the products are actually 0, so that $w = \mathbf{0}$. Necessarily, we thus have $\mathcal{K} = \{1, 3\}$ in the considered example.

Let us now check that the inequalities hold for any $k \in \mathcal{K}$. By definition of the data, we have for $k \in \{1, 3\}$:

$$y_k \|x_k\|^2 \geq (1 + (1-\eta)^2)^2.$$

And also

$$\begin{aligned} \sqrt{\sum_{k'=1}^{n} y_k^2} \sqrt{\sum_{k' \neq k} \langle x_{k'}, x_k \rangle} &\leq \sqrt{\eta^2 + (1+\eta)^2} \sqrt{(\eta(1-\eta) + (1+\eta)^2)^2 + (-(1-\eta)^2 + (1+\eta)^2)^2} \\ &= \sqrt{1 + 2\eta + 2\eta^2} \sqrt{(1+3\eta)^2 + 16\eta^2}. \end{aligned}$$

A simple comparison then allows to have when $\eta < \frac{1}{6}$:

$$y_k \|x_k\|^2 \geq (1 + (1-\eta)^2)^2 > \sqrt{\sum_{k'=1}^{n} y_k^2} \sqrt{\sum_{k' \neq k} \langle x_{k'}, x_k \rangle}.$$

So Assumption 5 holds.

3) Let's now check Assumption 6. Denote in the following for any $i \in [n]$, $r_i \coloneqq x_i^\top \beta^* - y_i$. By definition of β^*,

$$\sum_{i=1}^{3} r_i x_i = \mathbf{0}.$$

In particular, since x_1 and x_3 are linearly independent, it comes:

$$r_1 + \alpha_1 r_2 = 0 \quad \text{and} \quad r_3 + \alpha_3 r_2 = 0. \tag{13}$$

Also, by definition of r_i,

$$r_2 = \alpha_1 r_1 + \alpha_3 r_3 + \Big(\alpha_1 y_1 + \alpha_3 y_3 - y_2 \Big).$$

Since $\alpha_1 y_1 > y_2$ and $\alpha_3 > 0$, the term in parenthesis is positive, so that the last equality becomes

$$\begin{aligned} r_2 &> \alpha_1 r_1 + \alpha_3 r_3 \\ &= -(\alpha_1^2 + \alpha_3^2) r_2. \end{aligned}$$

The last inequality comes from Equation (13). Necessarily, this yields $r_2 > 0$, but Equation (13) then yields that $r_1 < 0$ and $r_3 < 0$, i.e., for any $k \in \mathcal{K}$, $y_k > \langle \beta^*, x_k \rangle$.

Now, consider $u \in A(\mathbb{R}^d)$ such that $\exists k, k'$ with $u_k = 1$ and $u_{k'} = -1$. Since $2 \notin \mathcal{K}$, we can actually choose k and k' both in $\mathcal{K}$. Assume without loss of generality that $u_1 = 1$ and $u_3 = -1$. There are now three cases, given by $u_2 \in \{-1, 0, 1\}$. If $u_2 = 0$, note that the considered cone is just given by a half line. So that for a small enough $\delta_u > 0$, there is no $k \in \mathcal{K}$ such that $\inf_{w \in A^{-1}(u)} |\langle \frac{w}{\|w\|}, \frac{x_k}{\|x_k\|} \rangle| \leq \delta_u$. The first point is then automatic in that case.

If instead $u_2 = 1$, then for a small enough δ_u, $k = 3$ is the only $k \in \mathcal{K}$ satisfying $\inf_{w \in A^{-1}(u)} |\langle \frac{w}{\|w\|}, \frac{x_k}{\|x_k\|} \rangle| \leq \delta_u$. From there, note that $\mathfrak{D}_u^{\beta^*} = \{\tilde{D}_u^{\beta^*}\}$ and

$$\begin{aligned} \tilde{D}_u^{\beta^*} &= \frac{1}{n} \sum_{k=1}^{2} (y_k - x_k^\top \beta^*) x_k \\ &= -\frac{1}{n} (y_3 - x_3^\top \beta^*) x_3. \end{aligned} \tag{14}$$

Since $(y_3 - x_3^\top \beta^*) > 0$, this indeed yields that $-\langle \tilde{D}_u^{\beta^*}, x_3 \rangle > 0$, which also yields the first point in that case.

The remaining case is $u_2 = -1$. In that case, we similarly only need to check it for $k = 1$ and

$$\tilde{D}_u^{\beta^*} = \frac{1}{n} (y_1 - x_1^\top \beta^*) x_1. \tag{15}$$

This then also yields the first point.

It now remains to check the second point of Assumption 6 in all 3 cases. The cases $u_2 = 0$ and $u_3 = -1$ are actually dealt together, since they have the same $\tilde{D}_u^{\beta^*}$. In that case, Equation (15) actually yields that

$$\begin{aligned} \inf_{w \in A_\delta^{-1}(u)} \langle \tilde{D}_u^{\beta^*}, \frac{w}{\|w\|} \rangle &= \frac{1}{n} (y_1 - x_1^\top \beta^*) \langle x_1, \frac{w}{\|w\|} \rangle \\ &\geq \frac{1}{n} (y_1 - x_1^\top \beta^*) \delta \|x_1\| > 0. \end{aligned}$$

A similar argument yields in the case $u_2 = 1$ to

$$\inf_{w \in A_\delta^{-1}(u)} \langle \tilde{D}_u^{\beta^*}, \frac{w}{\|w\|} \rangle \geq \frac{1}{n} (y_3 - x_3^\top \beta^*) \delta \|x_3\| > 0.$$

This concludes the proof. □

Appendix B. Additional Experiments

B.1 Experimental Details

In Section 5, we considered the following univariate 3 points dataset:

$$x_1 = -0.75 \text{ and } y_1 = 1.1;$$
$$x_2 = -0.5 \text{ and } y_2 = 0.1;$$
$$x_3 = 0.125 \text{ and } y_3 = 0.8.$$

The neural network architecture counts bias terms in the hidden layer, which we recall is equivalent to the parameterisation given by Equation (1), when adding 1 as the second coordinate to all the data points x_i. The activation function is ReLU, the initialisation follows Equation (3) with $\lambda = 10^{-3}$ and

$$\tilde{w}_j \sim \mathcal{N}(0, I_2),$$
$$\tilde{a}_j = s_j \|\tilde{w}_j\| \quad \text{with } s_j \sim \mathcal{U}(\{-1, 1\}).$$

We consider a **perfectly balanced** initialisation for three reasons:

- it is compatible with the initialisation regime described by Wojtowytsch (2020);
- it yields more interesting observations, since the neurons with $|a_j| > \|w_j\|$ actually align faster with extremal vectors;
- given the remark on the nature of the stationary point in Section 6, convergence towards a bad estimator is less obvious in this perfectly balanced case and might be more easily challenged by the use of finite step sizes.

Lastly, the neural network is trained with $m = 200\,000$ neurons to approximate the infinite width regime. We ran gradient descent with learning rate 10^{-3} up to 2 millions of iterations. The parameters clearly converged at this point and do not move anymore.

B.2 Stewart et al. Example

Figure 3 presents the training dynamics on another example of data. The considered example is here borrowed from Stewart et al. (2023). Precisely, we consider 40 univariate data points x_i sampled uniformly at random in $[-1, 1]$. The labels y_i are given by a teacher network closely resembling the one[8] of Stewart et al. (2023), so that $y_i = f^*(x_i)$ where f^* is a 8 neurons network.

The considered neural network follows the same parameterisation as given in Appendix B.1. We believe this is still a large enough width to approximate the infinite width regime here. The only other difference with Appendix B.1 is that the initialisation is here unbalanced; precisely, the weights $\tilde{w}_j$ and $\tilde{a}_j$ are drawn i.i.d. as

$$\tilde{w}_j \sim \mathcal{N}(0, I_2),$$
$$\tilde{a}_j \sim \mathcal{N}(0, 1).$$

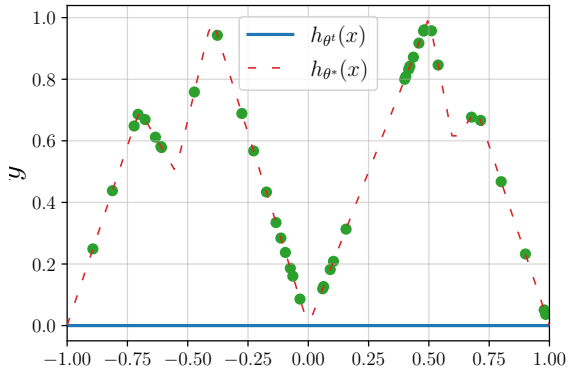
8. We just choose a simpler teacher network here by removing a small neuron in the representation.

This is an important remark as such an initialisation is closer to the initialisation chosen in practice and does not satisfy Equation (4), which is crucial to our analysis. Besides providing another, more complex example of convergence towards spurious stationary points, this experiment thus also illustrates that the balancedness condition given by Equation (4) might not be always necessary to yield convergence towards spurious stationary points.

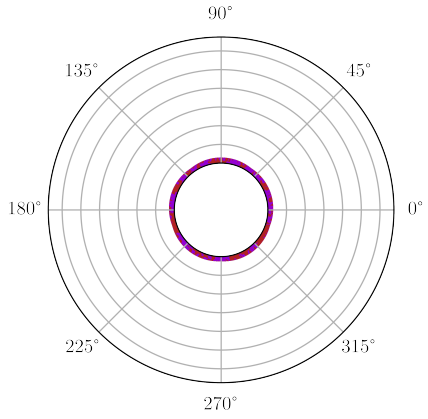
As opposed to Figure 2, the figures on the right column do not show the different activation cones and data points x_i here. This is just for the sake of clarity, since there are 40 points and 80 activation cones here.

Optimisation is hard on this example, as the teacher function counts two little bumps on the left and right slopes. These bumps are hard to learn while optimising, as they are restricted to small regions, i.e. activation cones. Since omnidirectionality is easily lost during training, having no remaining weights in this small region should be enough to get a "bad" training (i.e. convergence towards spurious stationary points).

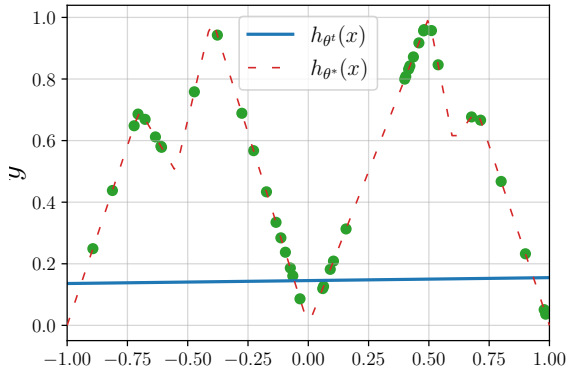
Figure 3 illustrates this difficulty: these small cones of interest indeed do not contain any remaining neuron after some point in the training. From there, it becomes impossible



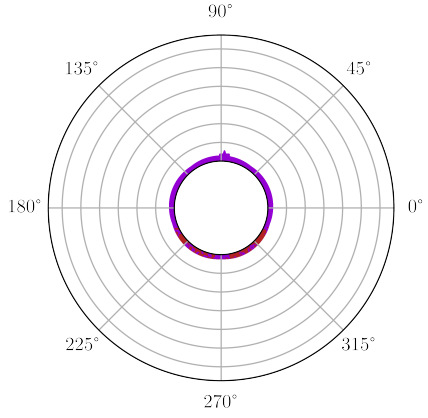
(a) Estimated function at initialisation (iteration 0)



(b) Weights' repartition at initialisation

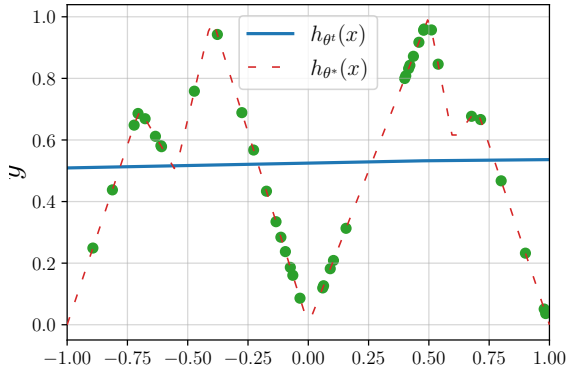


(c) Estimated function after early phase (iteration 6500)

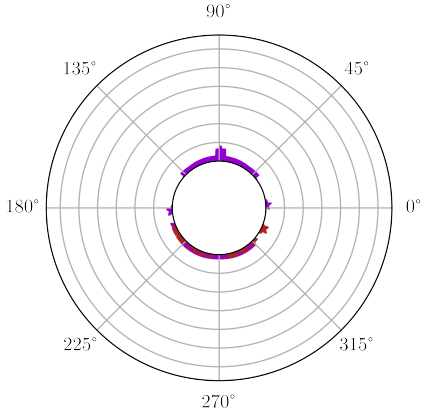


(d) Weights' repartition after early phase

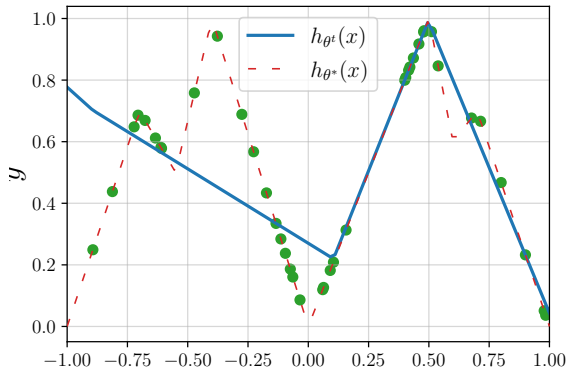
Figure 3: Training dynamics on Stewart et al. (2023) example (part 1/2).



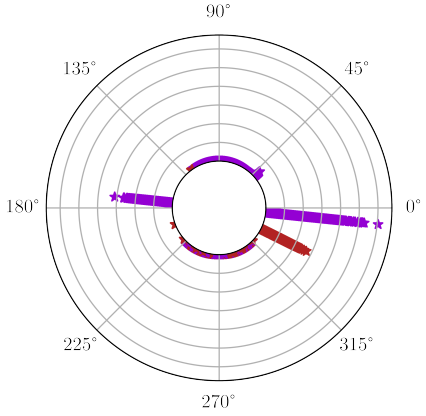
(e) Estimated function after first neuron growth (iteration 240 000)



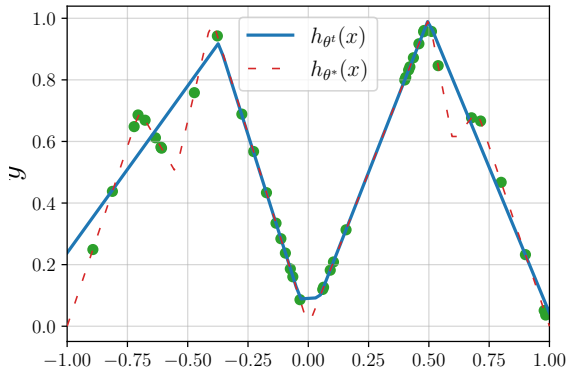
(f) Weights' repartition after first neuron growth



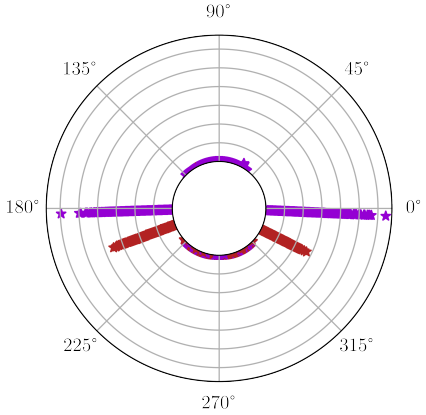
(g) Estimated function after second neuron growth (iteration 700 000)



(h) Weights' repartition after second neuron growth



(i) Estimated function at convergence (iteration 2×10^6)



(j) Weights' repartition at convergence

Figure 3: Training dynamics on Stewart et al. (2023) example (part 2/2).

to improve the current estimation, leading to convergence towards a spurious stationary point.

Figures 3a and 3b show the network at initialisation. Due to small scale of initialisation, both neurons and estimated function are nearly 0 here.

Figures 3c and 3d give the learnt parameters at the end of the early phase, just when some neurons start to considerably grow in norm and the estimated function is not zero anymore. A first interesting observation is that there still seems to be omnidirectionality of the weights here, as opposed to Figure 2d for the 3 points example. This is because the early alignment result of Theorem 1 only applies to neurons with $|a_j| \geq \|w_j\|$. The alignment of neurons with $|a_j| < \|w_j\|$ can happen at a much slower rate. Such neurons are thus not yet aligned with some extremal vector in Figure 3d.

Figures 3e and 3f show the state of the network after a first neurons growth. At this point, the estimated function is equivalent to a 1 neuron, constant, network. Also interestingly, Figure 3f highlights that omnidirectionality of the weights is already lost at this point of training. The neurons indeed seem to be concentrated in some specific cones, while other cones happen to have no neuron at all. This absence of neurons in key cones is at the origin of failure of training at convergence.

Figures 3g and 3h illustrates the learnt parameters after the second neurons growth. After this growth, the network is nearly equivalent to a 4 neurons network (a fifth one is starting to grow). Again, some activation cones do not include any neuron now.

Finally, Figures 3i and 3j show the state of the network at convergence. The estimated function is now equivalent to a 5 neurons network and thus fails at fitting all the data that is given by an 8 neurons network. The neurons that the network fails to learn correspond to the 2 little bumps mentioned above. At this point the network is unable to learn them, since the cones corresponding to these bumps are empty (of neurons). The network thus does not perfectly fit the data in the end. Yet, it gives a good 5 neurons approximation of the data; which confirms our conjecture made in Section 6 that despite imperfect data fitting, there still is some implicit bias incurred while training.

Appendix C. Proof of Intermediate Lemmas

C.1 Proof of Lemma 1

From Equation (6), it comes that a.e.

$$\begin{aligned}\frac{\mathrm{d}(a_i^t)^2}{\mathrm{d}t} - \frac{\mathrm{d}\|w_i^t\|^2}{\mathrm{d}t} &\in 2a_i^t\langle w_i^t, \mathfrak{D}_i^t\rangle - 2a_i^t\langle w_i^t, \mathfrak{D}_i^t\rangle \\ &= 0.\end{aligned}$$

The last equality comes from the fact that the scalar product does not depend on the choice of the subgradient $D \in \mathfrak{D}_i^t$. Finally, $(a_i^t)^2 - \|w_i^t\|^2$ is constant, which yields Lemma 1.

C.2 Proof of Lemma 2

By contradiction, assume that for some u such that $\mathbf{0} \notin \mathfrak{D}_u$, we have some $D \in \mathfrak{D}_u \cap \partial\overline{A^{-1}}(u)$. Let $(\eta_k)_{k\in[n]}$ in $[\gamma, 1]^n$ such that

$$D = -\frac{1}{n}\sum_{k=1}^{n} \eta_k \partial_1 \ell(0, y_k) x_k$$
$$\eta_k = 1 \text{ if } u_k = 1, \text{ and } \eta_k = \gamma \text{ if } u_k = -1.$$

Let $K(u) = \{k \in [n] \mid u_k = 0\}$. Also write in the following

$$\tilde{D}_u = -\frac{1}{n}\sum_{k=1}^{n} (\mathbb{1}_{u_k=1} + \gamma \mathbb{1}_{u_k=-1}) \partial_1 \ell(0, y_k) x_k.$$

Since $D \in \partial\overline{A^{-1}}(u)$, $\langle D, x_k\rangle = 0$ for any $k \in K(u)$. As a consequence, $D = P_{S(u)}(\tilde{D}_u)$, where $P_{S(u)}$ is the orthogonal projection on $\{x_k \mid k \in K(u)\}^\perp$. In particular, the previous argument allows to show the following

$$D'' \in \mathfrak{D}_u \text{ is extremal } \implies D'' = \operatorname*{argmin}_{D' \in \mathfrak{D}_u} \|D'\|. \tag{16}$$

Also, $D \in \partial\overline{A^{-1}}(u)$ implies there is at least one k such that $u_k \neq 0$ and $\langle D, x_k\rangle = 0$, i.e,

$$\partial_1 \ell(0, y_k)(\mathbb{1}_{u_k=1} + \gamma \mathbb{1}_{u_k=-1}) \|P_{S(u)}(x_k)\|^2 + \sum_{k' \neq k} \partial_1 \ell(0, y_{k'})(\mathbb{1}_{u_{k'}=1} + \gamma \mathbb{1}_{u_{k'}=-1}) \langle x_k, P_{S(u)}(x_{k'})\rangle = 0. \tag{17}$$

If at least one of the $\mathbb{1}_{u_{k'}=1} + \gamma \mathbb{1}_{u_{k'}=-1}$ is non-zero in the sum, then observe that conditionally on x_k and $\{x_{k'} \mid k' \in K(u)\}$ (all non-zero), this sum follows a continuous distribution. Hence, the event given by Equation (17) has 0 probability.

If instead, all the $\mathbb{1}_{u_{k'}=1} + \gamma \mathbb{1}_{u_{k'}=-1}$ are zero in the sum, then we necessarily have $\mathbb{1}_{u_k=1} + \gamma \mathbb{1}_{u_k=-1}$ non-zero, as we assumed that $\mathbf{0} \notin \mathfrak{D}_u$. In that case, Equation (17) becomes

$$\partial_1 \ell(0, y_k) \mathbb{1}_{u_k=1} \|P_{S(u)}(x_k)\|^2 = 0.$$

Again, the left term follows a continuous distribution, which leads to the first part of Lemma 2.

The last part is immediate given Assumption 2 and the fact that $\partial_1 \ell(0, y_k)$ is non-zero with probability 1, thanks to Assumption 1.

C.3 Proof of Lemma 3

Consider the maximisation program

$$\max_{w \in \mathbb{S}_d} G(w).$$

Since G is continuous and $\mathbb{S}_d$ is compact, the maximum is reached at some point w^*. The KKT conditions at w^* write

$$\mathbf{0} \in -\partial G(w^*) + \mu w^*,$$

for some $\mu \in \mathbb{R}$. Now note that $\partial G(w^*) = \mathfrak{D}(w^*, \mathbf{0})$, so that the KKT conditions rewrite

$$\mu w^* \in \mathfrak{D}(w^*, \mathbf{0}). \tag{18}$$

First assume $\mu \neq 0$. For any $v \in \mathfrak{D}(w^*, \mathbf{0})$, $\langle w^*, v\rangle$ has the same value. This means that any $v \in \mathfrak{D}(w^*, \mathbf{0})$ writes $v = \mu w^* + v^\perp$ where $v^\perp$ is orthogonal to w^*. As a consequence, $D(w^*, \mathbf{0}) = \mu w^*$ by norm minimisation. In particular, $A(D(w^*, \mathbf{0})) \in \{A(w^*), -A(w^*)\}$, so that $D(w^*, \mathbf{0})$ is extremal.

If instead $\mu = 0$, then $\mathbf{0} \in \mathfrak{D}(w^*, \mathbf{0})$, so that $D(w^*, \mathbf{0}) = 0$ and $G(w^*) = 0$. But then when considering the minimisation program

$$\min_{w \in \mathbb{S}_d} G(w),$$

the same reasoning yields that either there exists an extremal vector, or $\min_{w\in\mathbb{S}_d} G(w) = 0$. As a conclusion, if there is no extremal vector, G is constant equal to 0. But then, note that for any k

$$\begin{aligned} 0 = G(\frac{x_k}{\|x_k\|}) &= \frac{1}{n\|x_k\|}\sum_{k'=1}^n \partial_1\ell(0, y_{k'})\sigma(\langle x_k, x_{k'}\rangle) \\ &= \frac{1}{n\|x_k\|}\sum_{k'\neq k}^n \partial_1\ell(0, y_{k'})\sigma(\langle x_k, x_{k'}\rangle) + \frac{\partial_1\ell(0, y_k)}{n}\|x_k\|. \end{aligned}$$

If Assumption 2 holds, this equality only holds with probability 0, yielding Lemma 3.

Appendix D. Proof of Theorem 1

D.1 Additional Quantities

Define for any $u \in \{-1, 0, 1\}^n$,

$$\begin{gathered} Z_u \coloneqq \frac{1}{n}(\partial_1\ell(0, y_i)x_i)_{i\in K(u)} \in \mathbb{R}^{d\times|K(u)|}, \\ \text{where } K(u) \coloneqq \{k \in [n] \mid u_k = 0\}. \end{gathered} \tag{19}$$

In words, the columns of Z_u are given by the vectors $\frac{\partial_1\ell(0,y_i)x_i}{n}$ for all i such that $u_i = 0$. Let also in the following say that a set $\mathfrak{D}_u$ is **extremal** if there is some extremal vector $D \in \mathfrak{D}_u$, and define for any $u \in A(\mathbb{R}^d)$ and θ,

$$\begin{aligned} D_u^\theta &= D(w, \theta) \quad \text{ for some } w \in A^{-1}(u), \\ \tilde{D}_u &= -\frac{1}{n}\sum_{k=1}^n (\mathbb{1}_{u_k=1} + \gamma\mathbb{1}_{u_k=-1})\partial_1\ell(0, y_k)x_k. \end{aligned}$$

We shorten $D_u = D_u^{\mathbf{0}}$. Note that the definition does not depend on the choice of $w \in A^{-1}(u)$. We define

$$\begin{gathered} \delta_0 \coloneqq \min(\delta_0', \delta_0'') \\ \text{where } \delta_0' \coloneqq \min_{u, \mathfrak{D}_u \text{ is extremal}} \min_{k, u_k=0} \sigma_{\min}(Z_u)\min(\eta_k^u - \gamma, 1 - \eta_k^u) \\ \text{and } \delta_0'' \coloneqq \min_{u, \mathfrak{D}_u \text{ is extremal}} \min_{k, u_k\neq 0} \frac{|\langle D_u, x_k\rangle|}{\|x_k\|}. \end{gathered} \tag{20}$$

By convention in the following, we note $\min \emptyset = +\infty$. In Equation (20), $\sigma_{\min}(Z_u)$ is the smallest singular value of Z_u and $\boldsymbol{\eta}^u$ is the unique vector $\boldsymbol{\eta} \in \mathbb{R}^n$ satisfying[9] $D_u = \tilde{D}_u + Z_u \boldsymbol{\eta}$.

The value of $\lambda^*_{\alpha_0}$ in Theorem 1 is given by

$$\lambda^*_{\alpha_0} = \left(\min \left(\frac{n}{\sum_{k=1}^n \|x_k\|^2} \min(\frac{\alpha_{\min}^2}{8} D_{\min}, \frac{\alpha_0^2}{4} D_{\min}, \delta_0);\ \min_{k \in [n]} \frac{|\partial_1 \ell(0, y_k)|}{\|x_k\|} \right) \right)^{\frac{1}{2-4\varepsilon}}. \quad (21)$$

$$\text{where} \quad D_{\min} \coloneqq \min_{u, \mathbf{0} \notin \mathfrak{D}_u} \min_{D \in \mathfrak{D}_u} \|D\|;$$

$$\alpha_{\min} \coloneqq \min(\alpha_{\min,+}, \alpha_{\min,-}); \quad (22)$$

$$\alpha_{\min,+} \coloneqq \sqrt{1 - \left(\max_{\substack{u, \mathbf{0} \notin \mathfrak{D}_u \\ \mathfrak{D}_u \cap A^{-1}(u) = \emptyset}} \max_{w \in \overline{A^{-1}(u)} \setminus \{\mathbf{0}\}} \frac{\langle w, D_u \rangle}{\|w\| \|D_u\|} \right)^2} \quad (23)$$

$$\alpha_{\min,-} \coloneqq \sqrt{1 - \left(\min_{\substack{u, \mathbf{0} \notin \mathfrak{D}_u \\ \mathfrak{D}_u \cap -A^{-1}(u) = \emptyset}} \min_{w \in \overline{A^{-1}(u)} \setminus \{\mathbf{0}\}} \frac{\langle w, D_u \rangle}{\|w\| \|D_u\|} \right)^2} \quad (24)$$

and δ_0 is defined in Equation (20).

Lemmas 9 and 10 below imply that $\lambda^*_{\alpha_0}$ for any $\alpha_0 > 0$.

D.2 Additional Lemmas

Lemma 9. *Under Assumption 2, the quantities $\alpha_{\min}$ and $D_{\min}$ defined in Equation (22) almost surely satisfy $\alpha_{\min} > 0$ and $D_{\min} > 0$.*

Proof. First show that $\alpha_{\min,+} > 0$. The defined max is reached as the supremum of a continuous function on a compact set (the constraint set can indeed be made compact by restricting the problem to the sphere). By contradiction, if $\alpha_{\min,+} = 0$, since the max is reached, there is u such that $D_u \notin A^{-1}(u)$ and $\lambda D_u \in \overline{A^{-1}(u)}$ for some $\lambda > 0$. This implies that $D_u \in \partial \overline{A^{-1}(u)}$, which contradicts Lemma 2.

The same argument holds to show that $\alpha_{\min,-} > 0$, and thus $\alpha_{\min} > 0$

For $D_{\min}$, the infimum is also reached as each $\mathfrak{D}_u$ is a compact set. By definition, it is necessarily non-zero. □

Lemma 10. *Under Assumption 2, the quantity δ_0 defined in Equation (20) almost surely satisfies $\delta_0 > 0$.*

Proof. There is a finite number of u and k to consider in the min defined in δ_0' and δ_0'', so that we only need to prove that the quantity is positive for each u and k in the constraint set. We thus consider u and k in the constraint set in the following. Thanks to Equation (16), we necessarily have $\mathbf{0} \notin \mathfrak{D}_u$.

9. This vector is indeed unique when Z_u is injective, i.e., when $\sigma_{\min}(Z_u) > 0$. When it is not definitely unique, the value is then 0 as $\sigma_{\min}(Z_u) = 0$.

First, it is easy to show that $\delta_0'' > 0$. Thanks to Equation (16) again, D_u is extremal. In particular, for k such that $u_k \neq 0$, $\langle D_u, x_k\rangle \neq 0$, which implies that $\delta_0'' > 0$.

It is more technical for δ_0'. Since $u \neq \mathbf{0}$, $\mathrm{Span}\left((x_k)_{k\in K(u)}\right) \neq \mathbb{R}^d$, where $K(u) \coloneqq \{k \in [n] \mid u_k = 0\}$. Lemma 2 then implies that $|K(u)| < d$ and the vectors $(\partial_1\ell(0, y_i)x_k)_{k\in K(u)}$ are linearly independent. By definition of Z_u, this directly yields that Z_u is injective, i.e., $\sigma_{\min}(Z_u) > 0$ and η^u is indeed uniquely defined, and by definition of D_u, $\eta_k^u \in [\gamma, 1]$.

By contradiction, assume that $\eta_k^u = 1$ for some $k \in K(u)$. Then, we can define $\tilde{u} \in \{-1, 0, 1\}^n$

$$\tilde{u}_i = \begin{cases} 1 \text{ if } i = k \\ u_i \text{ if } i \neq k \end{cases}.$$

It then comes that $D_{\tilde{u}} = D_u$. In particular, $D_{\tilde{u}} \in -A^{-1}(u) \cup A^{-1}(u) \subset -\partial\overline{A^{-1}(\tilde{u})} \cup \partial\overline{A^{-1}(\tilde{u})}$. This directly contradicts Lemma 2, so that $\eta_k^u < 1$. A symmetric argument holds to prove that $\eta_k^u > \gamma$. In the end, we proved $\delta_0' > 0$, which allows to conclude. □

Lemma 11. *If Assumption 2 holds, then with probability* 1*, for any* θ *such that* $\frac{1}{n}\sum_{k=1}^n \|h_\theta(x_k)x_k\|_2 < \delta_0$ *and for any* $u \in A(\mathbb{R}^d)$*,*

1. $D_u = \mathbf{0} \implies D_u^\theta = \mathbf{0}$*;*
2. $D_u \in A^{-1}(u) \implies D_u^\theta \in A^{-1}(u)$*;*
3. $D_u \in -A^{-1}(u) \implies D_u^\theta \in -A^{-1}(u)$*.*

Proof. Let us prove the first two points of Lemma 11. The last part is proven similarly to the second point.

1) Consider some $u \in A(\mathbb{R}^d)$ such that $D_u = \mathbf{0}$ and define $K(u) \coloneqq \{k \in [n] \mid u_k = 0\}$. First recall that for some $\eta_k \in [\gamma, 1]$

$$D_u = -\sum_{k\notin K(u)} \partial_1\ell(0, y_k)x_k(\mathbb{1}_{u_k=1} + \gamma\mathbb{1}_{u_k=-1}) - \sum_{k\in K(u)} \eta_k\partial_1\ell(0, y_k)x_k$$

If $(\mathbb{1}_{u_k=1} + \gamma\mathbb{1}_{u_k=-1}) \neq 0$ for some $k \notin K(u)$, Lemma 2 implies with probability 1 that $|K(u)| \geq d$. This then implies $|K(u)| = n$, so that $(\mathbb{1}_{u_k=1} + \gamma\mathbb{1}_{u_k=-1}) = 0$ for any k. As a consequence, $D_u^\theta = \mathbf{0}$.

2) Assume that $D_u \in A^{-1}(u) \setminus \{\mathbf{0}\}$. By definition of δ_0'', for any $k \notin K(u)$, $|\langle D_u, x_k\rangle| \geq \delta_0\|x_k\|$. Moreover, note that by 1-Lipschitz property of $\partial_1\ell$,

$$\|D_u^\theta - D_u\| \leq \frac{1}{n}\sum_{k=1}^n \|h_\theta(x_k)x_k\|.$$

So that when $\frac{1}{n}\sum_{k=1}^n \|h_\theta(x_k)x_k\| < \delta_0$, $\langle D_u^\theta, x_k\rangle$ has the same sign than $\langle D_u, x_k\rangle$ for any $k \notin K(u)$.

It now remains to show for any $k \in K(u)$ that $\langle D_u^\theta, x_k\rangle = 0$. Let us write D_u as

$$D_u = \tilde{D}_u - \frac{1}{n}\sum_{k\in K(u)} \eta_k\partial_1\ell(0, y_k)x_k,$$

where $\tilde{D}_u \coloneqq -\frac{1}{n}\sum_{k=1}^n \partial_1\ell(0, y_k)x_k(\mathbb{1}_{u_k=1} + \gamma\mathbb{1}_{u_k=-1})$ and $\eta_k \in [\gamma, 1]$ for any $k \in K(u)$.

$D_u \in A^{-1}(u)$ implies by definition that $D_u \in \big((x_k)_{k\in K(u)}\big)^\perp$, so that $D_u = P_{S(u)}(\tilde{D}_u)$, where $P_{S(u)}$ is the orthogonal projection on the subspace $\big((x_k)_{k\in K(u)}\big)^\perp$.

Similarly, define $\tilde{D}_u^\theta \coloneqq -\frac{1}{n}\sum_{k=1}^n \partial_1\ell(\theta, y_k)x_k(\mathbb{1}_{u_k=1} + \gamma\mathbb{1}_{u_k=-1})$. We can also choose $(\eta'_k)_{k\in K(u)}\mathbb{R}^{K(u)}$ such that

$$P_{S(u)}(D(w,\theta)) - \tilde{D}(w,\theta) = -\frac{1}{n}\sum_{k\in K(u)} \eta'_k \partial_1\ell(\theta, y_k)x_k.$$

Note that the η'_k are not necessarily in $[\gamma, 1]$. Our goal is now to show they are actually in $[\gamma, 1]$. Denote for simplicity:

$$v = P_{S(u)}(\tilde{D}_u) - \tilde{D}_u$$
$$v^\theta = P_{S(u)}(\tilde{D}_u^\theta) - \tilde{D}_u^\theta.$$

Again, by Lipschitz property of the considered functions, it comes

$$\|v - v^\theta\| \le \frac{1}{n}\sum_{k=1}^n \|h_\theta(x_k)x_k\| < \delta_0.$$

Moreover, with the definition of Z_u given by Equation (19), note that

$$\begin{aligned}\|v - v^\theta\| &= \|Z_u(\eta - \eta')\| \\ &\ge \sigma_{\min}(Z_u)\|\eta - \eta'\|_2 \\ &\ge \sigma_{\min}(Z_u)\|\eta - \eta'\|_\infty.\end{aligned}$$

So overall, it comes

$$\|\eta - \eta'\|_\infty < \frac{1}{\sigma_{\min}(Z_u)}\delta_0.$$

Moreover, by definition of δ_0 (which is positive thanks to Lemma 10),

$$\min(\eta_k - \gamma, 1 - \eta_k) \ge \frac{\delta_0}{\sigma_{\min}(Z_u)} \quad \text{for any } k \in K(u).$$

This yields that $\eta'_k \in (\gamma, 1)$ for any $k \in K(u)$, i.e.

$$P_{S(u)}(\tilde{D}_u^\theta) \in \tilde{D}_u^\theta - \left\{ \frac{1}{n}\sum_{k\in S(u)} \zeta_k y_k x_k \mid \zeta_k \in (\gamma, 1) \text{ for any } k \in K(u) \right\}.$$

By minimization of the norm, this necessarily implies $D_u^\theta = P_{S(u)}(\tilde{D}_u^\theta)$. In particular, $\langle D_u^\theta, x_k\rangle = 0$ for any $k \in K(u)$. This finally yields that $A(D_u^\theta) = u$ and concludes the proof.

3) The last case is proved similarly to the second one. □

D.3 Proof of Theorem 1 (i)

We prove in this subsection the point (i) of Theorem 1.
Define the stopping time $t_1 = \min\left\{t \ge 0 \mid \sum_{j=1}^m (a_j^t)^2 \ge \lambda^{2-4\varepsilon}\right\}$. Thanks to Equation (5),

$t_1 > 0$. Moreover Equation (6) yields for any $t \le t_1$,

$$\begin{aligned}\left|\frac{\mathrm{d}a_j^t}{\mathrm{d}t}\right| &\le \|w_j^t\| \, \|D(w_j^t, \theta^t)\| \\ &\le \|w_j^t\| \left(D_{\max} + \frac{1}{n}\sum_{k=1}^n \|h_{\theta^t}(x_k)x_k\|\right) \\ &\le |a_j^t| \left(D_{\max} + \frac{\lambda^{2-4\varepsilon}}{n}\sum_{k=1}^n \|x_k\|^2\right).\end{aligned}$$

The first inequality comes from the fact that $\langle w_j^t, D\rangle$ is independent from the choice of $D \in \mathfrak{D}(w_j^t, \theta^t)$. The second one comes from observing that $\|D(w_j^t, \theta^t) - D(w_j^t, \mathbf{0})\| \le \frac{1}{n}\sum_{k=1}^n \|h_{\theta^t}(x_k)x_k\|$, thanks to the Lipschitz property of Assumption 1. Grönwall's inequality then implies for any $t \le \min(t_1, \tau)$:

$$|a_j^t| \le |a_j^0| \exp\left(t\left(D_{\max} + \frac{\lambda^{2-4\varepsilon}}{n}\sum_{k=1}^n \|x_k\|^2\right)\right).$$

Plugging the value of τ yields for $\lambda \le \lambda^*$,

$$|a_j^t| < |a_j^0|\lambda^{-\varepsilon\left(1+\frac{\lambda^{2-4\varepsilon}}{nD_{\max}}\sum_{k=1}^n \|x_k\|^2\right)} \le |a_j^0|\lambda^{-2\varepsilon} \quad \text{for any } t < \min(t_1, \tau).$$

This implies $t_1 \ge \tau$. Similarly, we have $|a_j^t| > |a_j^0|\lambda^{2\varepsilon}$ for any $t < \tau$, which yields the first point of Theorem 1.

The following sections aim at proving point (ii) of Theorem 1. In that objective, we first need to prove several technical intermediate lemmas that are essential to control the neurons trajectory.

D.4 Local Stability of Critical Manifolds

For any $u \in A(\mathbb{R}^d)$, we denote $\mathcal{M}_u \coloneqq A^{-1}(u)$ the activation manifold given by u. We also say a manifold $\mathcal{M}_u$ is **critical** if $D_u \in -\mathcal{M}_u \cup \mathcal{M}_u \cup \{\mathbf{0}\}$ and $\mathcal{M}_u \neq \{\mathbf{0}\}$, i.e. if it is associated to a critical direction of G. Similarly to the proof of Lemma 3, a study of the KKT points of G shows that some vector $v \in \mathcal{M}_u$ is a critical point[10] of G. Due to the absence of saddle points of G, such a point is actually a local extremum of G. This implies an even stronger stability property of the manifold $\mathcal{M}_u$. This observation is described precisely by Lemma 12 below.

Before that, define for any $v \in \mathbb{R}^d$,

$$\overline{D}_u(v) = \lim_{t \to 0^+} D(w + tv, \mathbf{0}) \quad \text{for any } w \in A^{-1}(u).$$

The definition of $\overline{D}_u(v)$ is valid as it does not depend on $w \in A^{-1}(u)$. More precisely, we have $\overline{D}_u(v) = D_{u(v)}$, where

$$u(v)_k = \begin{cases} u_k \text{ if } u_k \neq 0 \\ \operatorname{sign}(\langle v, x_k\rangle) \text{ otherwise.}\end{cases}$$

10. v is of the form $\pm\frac{D_u}{\|D_u\|}$ if $D_u \neq \mathbf{0}$.

Lemma 12. *Assume G does not have any saddle point, Assumption 2 holds, θ satisfies the condition of Lemma 11 and for any $k \in [n]$, $|h_\theta(x_k)| < |\partial_1 \ell(0, y_k)|$. Then with probability 1, for any critical manifold $\mathcal{M}$, $\exists \varepsilon_{\mathcal{M}} \in \{-1, 1\}$, such that $\forall v \in \overline{\mathcal{M}}^\perp$,*

$$\varepsilon_{\mathcal{M}} \langle v, D^\theta_{u(v)} \rangle \leq 0.$$

Lemma 12 states that for any critical manifold $\mathcal{M}$, the gradients of G around $\overline{\mathcal{M}}$ either all points toward $\overline{\mathcal{M}}$, or point outside $\overline{\mathcal{M}}$. The case $\varepsilon_{\mathcal{M}} = 1$ corresponds to the former case, in which case a local maximum of G lies in $\mathcal{M}$. In that case, Lemma 12 implies that the manifold $\overline{\mathcal{M}}$ is locally stable when running gradient ascent over the function G, even with a small perturbation in the function G due to the parameters θ. Conversely, $\varepsilon_{\mathcal{M}} = -1$ corresponds to a local minimum, that is stable when running gradient descent over G.

Proof of Lemma 12. Let $\mathcal{M} = A^{-1}(u)$ be a critical manifold. Necessarily, there is $\overline{w} \in \mathbb{S}_d \cap \mathcal{M}$ a critical point of G. By assumption it is either a local minimum or maximum of G. Without loss of generality, we assume in the following it is a local maximum. The case of local minimum is dealt with similarly, after a change of sign for $\varepsilon_{\mathcal{M}}$.

Let us first show that for any $v \in \overline{w}^\perp$, $\langle v, D_{u(v)} \rangle \leq 0$. Let $v \in \mathbb{S}_d \cap \overline{w}^\perp$ and define for any $\delta > 0$,

$$w(\delta) = \sqrt{1 - \delta^2}\overline{w} + \delta v \in \mathbb{S}_d.$$

Since $\overline{w}$ is a local maximum of G, it comes for small enough δ that $G(w) \geq G(w(\delta))$. Moreover, as $D(w(\delta), \mathbf{0})$ is piecewise constant, it comes for small enough δ by definition of $\overline{D}_u(v)$ that

$$G(w(\delta)) = \langle w(\delta), \overline{D}_u(v) \rangle.$$

Moreover, by continuity of G it comes $G(\overline{w}) = \langle \overline{w}, \overline{D}_u(v) \rangle$, so that

$$0 \geq \langle w(\delta) - \overline{w}, D_{u(v)} \rangle.$$

Noting that $w(\delta) - \overline{w} = \delta v + \mathcal{O}\left(\delta^2\right)$, this necessarily implies that $\langle v, D_{u(v)} \rangle \leq 0$. By a rescaling argument, we thus have for any $v \in \overline{w}^\perp$,

$$\langle v, D_{u(v)} \rangle \leq 0. \tag{25}$$

Now assume for $v \in \overline{w}^\perp$ that $\langle v, D_{u(v)} \rangle = 0$. Let $S(u(v)) = \mathrm{Span}\{x_k \mid u(v)_k = 0\}$. Note that for any $\Delta \in S(u(v))^\perp$, $D_{u(v)} = D_{u(v+t\Delta)}$ for a small enough $t > 0$. By Equation (25), we then have for any $\Delta \in (S(u(v)) \cup \{\overline{w}\})^\perp$,

$$\langle \Delta, D_{u(v)} \rangle \leq 0,$$

which actually implies $\langle \Delta, D_{u(v)} \rangle = 0$ by symmetry, i.e.

$$D_{u(v)} \in S(u(v)) + \mathbb{R}\overline{w}.$$

Necessarily, $D_{u(v)} \in S(u(v)) + G(\overline{w})\overline{w}$ since $\langle D_{u(v)}, \overline{w} \rangle = G(\overline{w})$. The KKT conditions at $\overline{w}$ imply that $D_u \in \mathbb{R}\overline{w}$ and again, this implies that $D_u = G(w)\overline{w}$. From there, we then have α_k such that

$$D_{u(v)} = D_u + \sum_{k, u(v)_k = 0} \alpha_k x_k.$$

But by definition of D_u and $D_{u(v)}$, we also have η_k and η'_k such that

$$D_{u(v)} = D_u - \frac{1}{n} \sum_{k,u_k=0} (\eta'_k - \eta_k) \partial_1 \ell(0, y_k) x_k. \tag{26}$$

In particular, both equations imply

$$\sum_{k,u_k=0} \left(\alpha_k \mathbb{1}_{\langle v, x_k \rangle = 0} + \frac{\eta'_k - \eta_k}{n} \partial \ell_1(0, y_k) \right) x_k = 0.$$

Lemma 2 then implies a.s. that either $u_k = 0$ for at least d values of k; or $\alpha_k \mathbb{1}_{\langle v, x_k \rangle = 0} + \frac{\eta'_k - \eta_k}{n} \partial \ell_1(0, y_k) = 0$ for any k such that $u_k = 0$.

The former yields that $S(u) = \mathbb{R}^d$. Since $\overline{w} \in S(u)^\perp$, this yields $\overline{w} = 0$, which contradicts $\overline{w} \in \mathbb{S}_d$. Necessarily, $\alpha_k \mathbb{1}_{\langle v, x_k \rangle = 0} + \frac{\eta'_k - \eta_k}{n} \partial \ell_1(0, y_k) = 0$ for any k such that $u_k = 0$. In particular, for any k such that $u(v)_k \neq 0$, $\eta_k = \eta'_k$. Then note that Equation (26) becomes

$$D_{u(v)} = D_u - \frac{1}{n} \sum_{k,u(v)_k=0} (\eta'_k - \eta_k) \partial_1 \ell(0, y_k) x_k.$$

For any k such that $u(v)_k = 0$, η'_k is chosen in $[\gamma, 1]$ to minimise the norm of $D_{u(v)}$. The choice $\eta'_k = \eta_k$ then minimises the norm, so that $D_{u(v)} = D_u$. In that case Lemma 2 necessarily implies that $D_{u(v)} = \mathbf{0}$ (otherwise, we would have $D_{u(v)} = D_u \in -\mathcal{M} \cup \mathcal{M} \subset -\partial \overline{A^{-1}(u(v))} \cup \partial \overline{A^{-1}(u(v))}$).

We just showed that for any $v \in \overline{w}^\perp$

$$\langle v, D_{u(v)} \rangle < 0 \quad \text{or} \quad D_{u(v)} = \mathbf{0}. \tag{27}$$

Now consider any $v \in \overline{w}^\perp$. First, if $D_{u(v)} = \mathbf{0}$, then Lemma 11 directly implies that $D^\theta_{u(v)} = \mathbf{0}$. Now assume that $D_{u(v)} \neq \mathbf{0}$, so that $\langle D_{u(v)}, v' \rangle < 0$ for any $v' \in \mathcal{S}_v$ where

$$\mathcal{S}_v \coloneqq \{v' \in \overline{w}^\perp \mid u(v') = u(v)\}.$$

Let $i \in K(u)$ in the following. First note that the set $\mathrm{Span}(\{x_k \mid k \in K(u), k \neq i\})$ is of dimension at most $d-2$ with probability 1, since $\dim(\mathrm{Span}(\{x_k \mid k \in K(u)\})) \leq d-1$. Indeed, if this last set was $\mathbb{R}^d$, we would have $D_u = \mathbf{0}$. In particular, $(\{x_k \mid k \in K(u), k \neq i\} \cup \{\overline{w}\})^\perp$ is of dimension at least 1. Moreover, Assumption 2 along with $\overline{w} \perp K(u)$ implies that a.s.

$$x_i \notin \mathrm{Span}(\{x_k \mid k \in K(u), k \neq i\}) + \mathbb{R}\overline{w}.$$

This directly implies that

$$(\{x_k \mid k \in K(u), k \neq i\} \cup \{\overline{w}\})^\perp \not\subset \{x_i\}^\perp.$$

Therefore, there is some $v_i \in (\{x_k \mid k \in K(u), k \neq i\} \cup \{\overline{w}\})^\perp$ such that $\langle v_i, x_i \rangle = 1$. Now note that

$$\begin{aligned} 0 &\geq \langle v_i, D_{u(v_i)} \rangle \\ &= -\frac{\eta'_i - \eta_i}{n} \partial \ell_1(0, y_i), \end{aligned} \tag{28}$$

where

$$D_u = -\frac{1}{n}\sum_{k=1}^{n} \eta_k \partial \ell_1(0, y_k) x_k$$
$$D_{u(v_i)} = -\frac{1}{n}\sum_{k=1}^{n} \eta'_k \partial \ell_1(0, y_k) x_k.$$

Equation (28) comes from the fact that η_k and η'_k coincide for any $k \notin K(u)$; for $k \in K(u)$, we then used the fact that $\langle v_i, x_k \rangle = \mathbb{1}_{k=i}$.

In particular, Lemma 10 implies that $\eta_i \in (\gamma, 1)$, so that $\eta'_i - \eta_i = 1 - \eta_i > 0$. Equation (28) then implies that $\partial \ell_1(0, y_i) \geq 0$. This then holds for any $i \in K(u)$. By Assumptions 1 and 2 the inequality is actually strict. In particular, for any θ satisfying the inequality in Lemma 12, it also comes $\partial \ell_1(h_\theta(x_i), y_i) > 0$. Moreover, Lemma 11 also yields that $D^\theta_u \in -\mathcal{M} \cup \mathcal{M}$. This then yields for any $v \in \overline{\mathcal{M}}^\perp$:

$$\langle v, D^\theta_{u(v)} \rangle = -\frac{1}{n} \sum_{\substack{k \in K(u) \\ u(v)_k \neq 0}} (\eta'_k(\theta) - \eta_k(\theta)) \partial_1 \ell(h_\theta(x_k), y_k) \langle v, x_k \rangle.$$

Note that $\eta'_k(\theta) - \eta_k(\theta)$ is of the same sign than $\langle v, x_k \rangle$ (or zero). So that finally, every summand is non-negative.

Finally, we just showed that for any $v \in \overline{\mathcal{M}}^\perp$ and θ with h_θ small enough,

$$\langle v, D^\theta_{u(v)} \rangle \leq 0.$$

□

D.5 Global Stability

Lemma 13. *If Assumption 2 holds, the function G does not admit any saddle point and $\lambda < \lambda_1^*$, then for any $j \in [m]$ and $t \in [0, \tau]$,*

$$a_j^t D(w_j^t, \theta^t) \in A^{-1}(A(w_j^t)) \cup \{\mathbf{0}\} \text{ and } w_j^t \neq \mathbf{0} \implies w_j^{t'} \in A^{-1}(A(w_j^t)) \text{ for any } t' \in [t, \tau].$$

Proof. Recall that λ_1^* is given by Equation (21) for $\alpha_0 = 1$. The first point of Theorem 1 holds thanks to Appendix D.3. Thus, for any $t \leq \tau$, θ^t satisfies the conditions of Lemma 12, thanks to our choice of λ. Let $j \in [m]$ and $t_0 \in [0, \tau]$ such that $a_j^{t_0} D(w_j^{t_0}, \theta^{t_0}) \in A^{-1}(A(w_j^{t_0})) \cup \{\mathbf{0}\}$ and $w_j^t \neq \mathbf{0}$. For $u \coloneqq A(w_j^{t_0})$, note that $\mathcal{M} = \mathcal{M}_u$ is a critical manifold by definition. Assume in the following that $a_j^{t_0} > 0$. The negative case is dealt with similarly.

The definition of $\alpha_{\min}$ implies if D_u is not extremal that

$$\min_{w \in \overline{A^{-1}(u)}} 1 - \frac{\langle w, D_u \rangle^2}{\|w\|^2 \|D_u\|^2} \geq \alpha_{\min}^2.$$

First, note that when $\frac{1}{n}\sum_{k=1}^n \|h_\theta(x_k) x_k\| \leq \delta$,

$$\left\| \frac{D^\theta_u}{\|D^\theta_u\|} - \frac{D_u}{\|D_u\|} \right\| \leq \frac{2\delta}{\|D_u\| - \delta}$$

As a consequence, our choice of λ would imply if[11] D_u was not extremal that

$$\min_{w \in \overline{A^{-1}(u)}} 1 - \frac{\langle w, D_u^\theta \rangle^2}{\|w\|^2 \|D_u^\theta\|^2} > 0,$$

which contradicts that $D(w_j^{t_0}, \theta^{t_0}) \in A^{-1}(A(w_j^{t_0})) \cup \{\mathbf{0}\}$. Necessarily, for our choice of λ and thanks to Lemma 11,

$$\begin{aligned} D_u^\theta \in A^{-1}(u) &\implies D_u \in A^{-1}(u) \\ &\implies D_u^{\theta'} \in A^{-1}(u). \end{aligned}$$

and similarly with $D_u^\theta \in -A^{-1}(u)$ and $D_u^\theta = \mathbf{0}$. As a consequence, for any $t \in [t_0, \tau]$, it comes that $a_j^t D(w_j^{t_0}, \theta^t) \in A^{-1}(A(w_j^{t_0})) \cup \{\mathbf{0}\}$.

If $t_0 = 0$, then with probability 1, $\mathcal{M}$ is an open manifold of dimension d, i.e. $u_k \neq 0$ for all k. The inclusion $a_j^t D(w_j^{t_0}, \theta^t) \in A^{-1}(A(w_j^{t_0})) \cup \{\mathbf{0}\}$ then directly implies that the value of $|\langle x_k, w_j^t \rangle|$ is increasing over time on $[t_0, \tau]$ as

$$\frac{\mathrm{d}\langle x_k, w_j^t \rangle}{\mathrm{d}t} = a_j^t \langle x_k, D(w_j^{t_0}, \theta^t) \rangle.$$

Thus for any $t \in [t_0, \tau]$, $w_j^t \in A^{-1}(A(w_j^{t_0}) \cup \{\mathbf{0}\}$.

Now assume that we cannot choose $t_0 = 0$, i.e. $a_j^0 D(w_j^0, \theta^0) \notin A^{-1}(A(w_j^0)) \cup \{\mathbf{0}\}$. The distance of w_j^t to $\overline{\mathcal{M}}$ evolves as follows a.e., with $P_{\overline{\mathcal{M}}}$ the orthogonal projection on $\mathrm{Span}(\overline{\mathcal{M}})$,

$$\begin{aligned} \frac{\mathrm{d}d(w_j^t, \mathrm{Span}(\overline{\mathcal{M}}))}{\mathrm{d}t} &= 2\langle \frac{\mathrm{d}(w_j^t - P_{\overline{\mathcal{M}}}(w_j^t))}{\mathrm{d}t}, w_j^t - P_{\overline{\mathcal{M}}}(w_j^t) \rangle \\ &= 2\langle \frac{\mathrm{d}w_j^t}{\mathrm{d}t}, w_j^t - P_{\overline{\mathcal{M}}}(w_j^t) \rangle \end{aligned}$$

The second equality comes from the fact that $\frac{\mathrm{d}P_{\overline{\mathcal{M}}}(w_j^t)}{\mathrm{d}t} \in \mathrm{Span}(\overline{\mathcal{M}})$, while $w_j^t - P_{\overline{\mathcal{M}}}(w_j^t) \in \overline{\mathcal{M}}^\perp$.

Since $w_j^{t_0} \in \mathcal{M}$ and w_j^t is continuous in time, there exists $\eta > 0$ such that for any $t \in [t_0 - \eta, t_0 + \eta]$, $\{k \mid \langle x_k, w_j^t \rangle = 0\} \subset \{k \mid u_k = 0\}$, where we recall $\mathcal{M} = \mathcal{M}_u$. As a consequence, $\frac{\mathrm{d}w_j^t}{\mathrm{d}t} \in D(w_j^t, \theta^t) + \mathrm{Span}(\{x_k \mid \langle x_k, w_j^t \rangle = 0\}) \subset D(w_j^t, \theta^t) + \overline{\mathcal{M}}^\perp \cap \{w_j^t\}^\perp$. This then implies

$$\begin{aligned} \frac{\mathrm{d}d(w_j^t, \mathrm{Span}(\overline{\mathcal{M}}))}{\mathrm{d}t} &= 2\langle \frac{\mathrm{d}w_j^t}{\mathrm{d}t}, w_j^t - P_{\overline{\mathcal{M}}}(w_j^t) \rangle \\ &= 2\langle D(w_j^t, \theta^t), w_j^t - P_{\overline{\mathcal{M}}}(w_j^t) \rangle. \end{aligned}$$

Moreover by continuity, we can also choose $\eta > 0$ such that for any $t \in [t_0 - \eta, t_0 + \eta]$, $D(w_j^t, \theta^t) = D_{u(v^t)}^{\theta^t}$ where $v^t = w_j^t - P_{\overline{\mathcal{M}}}(w_j^t) \in \overline{\mathcal{M}}^\perp$. Lemma 12 then implies a.s. for any $t \in [t_0 - \eta, t_0 + \eta]$:

$$\varepsilon_{\mathcal{M}} \frac{\mathrm{d}d(w_j^t, \mathrm{Span}(\overline{\mathcal{M}}))}{\mathrm{d}t} \leq 0. \tag{29}$$

11. We here use $\frac{1}{n}\sum_{k=1}^n \|x_k\|^2 \lambda^{2-4\varepsilon} \leq \frac{1}{8} D_{\min} \alpha_{\min}^2$.

We can choose $\eta > 0$ so that $w_j^{t_0-\eta} \notin \mathrm{Span}(\overline{\mathcal{M}})$. In that case, Equation (29) implies that[12] $\varepsilon_{\mathcal{M}} = 1$. The $\varepsilon_{\mathcal{M}} = -1$ case would indeed have resulted in a growth of $d(w_j^t, \mathrm{Span}(\overline{\mathcal{M}}))$ on $[t_0 - \eta, t_0]$, contradicting the fact that it is non zero at $t_0 - \eta$ and zero at t_0.

As a consequence, for any $t_1 \leq \tau$ such that $w_j^{t_1}$, we have for similar reasons that $d(w_j^t, \mathrm{Span}(\overline{\mathcal{M}}))$ is non-increasing on $[t_1, t_1+\eta]$, so that $w_j^t \in \mathrm{Span}(\overline{\mathcal{M}})$ for any $t \in [t_1, t_1+\eta]$ for a small enough η. Moreover, note that $a_j^t D(w_j^{t_0}, \theta^t) \in A^{-1}(A(w_j^{t_0})) \cup \{\mathbf{0}\}$ implies here again that the values of $|\langle x_k, w_j^t \rangle|$ are non-decreasing for any k as long as $w_j^t \in \mathcal{M}$ and $t \leq \tau$. The last two points thus imply that the conditions $\langle w_j^t, x_k \rangle = 0$ are stable and the conditions $u_k \langle w_j^t, x_k \rangle > 0$ are enforced as long as $w_j^t \in \mathcal{M}$ and $t \leq \tau$. This concludes the proof of Lemma 13. □

D.6 Quantization of Misalignment

Lemma 14. *For any critical manifold $\mathcal{M}_u$ such that $D_u \neq \mathbf{0}$, if $w_u \in \mathcal{M}_u \cap \mathbb{S}_d$ is a local maximum of G, then either $D_u \in \mathcal{M}_u$ or $\mathcal{M}_u \cap \mathbb{S}_d = \{-\frac{D_u}{\|D_u\|}\}$.*

Similarly for any critical manifold $\mathcal{M}_u$ such that $D_u \neq \mathbf{0}$, if $w_u \in \mathcal{M}_u \cap \mathbb{S}_d$ is a local minimum of G, then either $D_u \in -\mathcal{M}_u$ or $\mathcal{M}_u \cap \mathbb{S}_d = \{\frac{D_u}{\|D_u\|}\}$.

Proof. Consider the first case of Lemma 14. The KKT condition of Equation (18) at w_u then yields that $\mu w_u = D_u$. Since $D_u \neq \mathbf{0}$, this necessarily yields that $w_u = \pm \frac{D_u}{\|D_u\|}$. If $w_u = \frac{D_u}{\|D_u\|}$, we then simply have $D_u \in \mathcal{M}_u$. Now assume that $w_u = -\frac{D_u}{\|D_u\|}$. Note that w_u is a local maximum of G and $G(w_u) = -\|D_u\|$ in that case. Now observe that if $\mathcal{M}_u \cap \mathbb{S}_d \neq \{-\frac{D_u}{\|D_u\|}\}$, we can choose $w \in \mathcal{M}_u \cap \mathbb{S}_d$ that is not perfectly aligned with $-D_u$. In particular

$$G(w) = \langle w, D_u \rangle > -\|D_u\|.$$

Necessarily, this yields that $\mathcal{M}_u \cap \mathbb{S}_d = \{-\frac{D_u}{\|D_u\|}\}$.

The second point of Lemma 14 is proved with symmetric arguments. □

D.7 Proof of Theorem 1 (ii)

Define for the remaining of the proof $\mathsf{w}_j^t = \frac{w_j^t}{a_j^t}$ and assume both $a_j^0 > 0$ and j satisfies Condition 2. We then have $a_j^t > 0$ for any t thanks to Lemma 1 (and a simple Grönwall argument if $\|w_j^0\| = |a_j^0|$). The symmetric case ($a_j^0 < 0$) is dealt with similarly. Note that $\mathsf{w}_j^t \in B(0,1)$. Moreover, we have for any $j \in [m]$ the differential inclusion

$$\frac{\mathrm{d}\mathsf{w}_j^t}{\mathrm{d}t} \in \mathfrak{D}_j^t - \langle \mathsf{w}_j^t, \mathfrak{D}_j^t \rangle \mathsf{w}_j^t. \tag{30}$$

Here again, the set $\langle \mathsf{w}_j^t, \mathfrak{D}(w_j^t, \theta) \rangle$ is a singleton for any θ, i.e. the scalar product does not depend on the choice of the vector $D \in \mathfrak{D}(w_j^t, \theta)$, so that Equation (30) rewrites for any $D, D' \in \mathfrak{D}(w_j^t, \mathbf{0})$:

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D \rangle}{\mathrm{d}t} = \frac{\mathrm{d}\langle \mathsf{w}_j^t, D' \rangle}{\mathrm{d}t} \in \langle \mathfrak{D}_j^t, D' \rangle - \langle \mathsf{w}_j^t, D(w_j^t, \theta^t) \rangle \langle \mathsf{w}_j^t, D \rangle.$$

12. The symmetric case $a_j^t < 0$ would here result in $\varepsilon_{\mathcal{M}} = -1$.

$D(w_j^t, \mathbf{0})$ is piecewise constant, so that $\frac{\mathrm{d}D(w_j^t,\mathbf{0})}{\mathrm{d}t} = 0$ almost everywhere. Also, the quantity $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ is absolutely continuous. In particular, for almost any $t \leq \tau$ and $D' \in \mathfrak{D}(w_j^t, \mathbf{0})$,

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} \in \langle D', \mathfrak{D}_j^t\rangle - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle\langle \mathsf{w}_j^t, D(w_j^t, \theta^t)\rangle.$$

Note that for any $D \in \mathfrak{D}_j^t$, there exists $D' \in \mathfrak{D}(w_j^t, \mathbf{0})$ such that $\|D - D'\| \leq \frac{1}{n}\sum_{k=1}^n \|h_{\theta^t}(x_k)x_k\|$, which yields

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} &\geq \max_{D' \in \mathfrak{D}(w_j^t,\mathbf{0})} \min_{D \in \mathfrak{D}_j^t} \langle D, D'\rangle - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle\langle \mathsf{w}_j^t, D(w_j^t, \theta^t)\rangle \\ &\geq \min_{D' \in \mathfrak{D}(w_j^t,\mathbf{0})} \left(\|D'\|^2 - \frac{\|D'\|}{n}\sum_{k=1}^n \|h_{\theta^t}(x_k)x_k\|\right) - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle\langle \mathsf{w}_j^t, D(w_j^t, \theta^t)\rangle.\end{aligned}$$

Moreover, thanks to the first part of Theorem 1

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} &\geq \min_{D' \in \mathfrak{D}(w_j^t,\mathbf{0})} \left(\|D'\|^2 - \frac{\|D'\|}{n}\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2\right) \\ &\quad - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 - \frac{\|D(w_j^t, \mathbf{0})\|}{n}\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2 \\ &\geq \|D(w_j^t, \mathbf{0})\|^2 - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 - \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2. \qquad (31)\end{aligned}$$

The last inequality comes from $\lambda < \lambda^*_{\alpha_0}$, which yields that the minimum is reached for the minimal value of $\|D'\|$ in the considered set.

Let $u_j^t = A(w_j^t)$ and $D_{u_j^t} \coloneqq D(w_j^t, \mathbf{0}) \in \mathfrak{D}_{u_j^t}$ in the following. If $D_{u_j^t} \notin A^{-1}(u_j^t) \cup \{\mathbf{0}\}$, the definition of $\alpha_{\min,+}$ (which is positive thanks to Lemma 9), given by Equation (22), implies that

$$\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle \leq \|D(w_j^t, \mathbf{0})\|\sqrt{1 - \alpha_{\min}^2}.$$

Equation (31) implies that as long as $\|D(w_j^t, \mathbf{0})\|^2 - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 \geq \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2$, the quantity $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ is increasing. In particular, the choice of $\lambda^*_{\alpha_0}$ makes it increasing as long as $D_{u_j^t} \notin A^{-1}(u_j^t) \cup \{\mathbf{0}\}$ and $\langle \mathsf{w}_j^t, D_{u_j^t}\rangle \geq -\sqrt{1 - \min(\alpha_{\min}, \alpha_0)^2}\|D(w_j^t, \mathbf{0}\|$. Note that this inequality holds at initialisation. Moreover when entering a new activation cone, either $\langle \mathsf{w}_j^t, D_{u_j^t}\rangle \geq -\sqrt{1 - \min(\alpha_{\min,-}, \alpha_0)^2}\|D_{u_j^t}\|$ or $D_{u_j^t} \in -A^{-1}(u_j^t) \cup \{\mathbf{0}\}$. We thus have by continuous induction that as long as $D_{u_j^t} \notin A^{-1}(u_j^t) \cup \{\mathbf{0}\}$ and w_j^t does not **enter** an activation cone such that $D_{u_j^t} \in -A^{-1}(u_j^t)$:

$$-\|D(w_j^t, \mathbf{0})\|\sqrt{1 - \min(\alpha_{\min}, \alpha_0)^2} \leq \langle \mathsf{w}_j^t, D_{u_j^t}\rangle \leq \|D(w_j^t, \mathbf{0})\|\sqrt{1 - \alpha_{\min}^2}.$$

Equation (31) and our choice of λ then imply that as long as $D_{u_j^t} \notin A^{-1}(u_j^t) \cup \{\mathbf{0}\}$ and w_j^t does not **enter** an activation cone such that $D_{u_j^t} \in -A^{-1}(u_j^t)$:

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} \geq \frac{D_{\min}^2}{2}\min(\alpha_{\min}, \alpha_0)^2.$$

Since $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ is bounded in absolute value by $D_{\max}$, there is a time t_2 bounded as $t_2 \leq \frac{4D_{\max}}{D_{\min}^2 \min(\alpha_{\min},\alpha_0)^2} = \Theta_{\alpha_0}(1)$ such that

$$D_{u_j^{t_2}} \in A^{-1}(u_j^{t_2}) \cup \{\mathbf{0}\} \tag{32}$$

$$\text{or } D_{u_j^{t_2}} \in -A^{-1}(u_j^{t_2}) \text{ and } u_j^{t_2} \neq u_j^0. \tag{33}$$

Note that either $\lambda^\varepsilon = \Omega_{\alpha_0}(1)$ or $t_2 \leq \tau$. The first case trivially yields the second point of Theorem 1, so that we assume from now that $t_2 \leq \tau$.

In the case of Equation (33), an argument similar to the proof of Lemma 13 yields for $\mathcal{M} = A^{-1}(u_j^{t_2})$ that $\varepsilon_{\mathcal{M}} = 1$, where $\varepsilon_{\mathcal{M}}$ is defined by Lemma 12. This corresponds to the case where a local maximum of G lies in $\mathcal{M}$. But since $D_{u_j^{t_2}} \in -A^{-1}(u_j^{t_2})$, Lemma 14 implies that $\mathcal{M} = \mathbb{R}_-^* D_{u_j^{t_2}}$.

Similarly to the proof of Lemma 13, the conditions $\langle w_j^t, x_k\rangle = 0$ are then stable as long as $w_j^t \in A^{-1}(u_j^{t_2})$. But since w_j^t is proportional to $D_{u_j^t}$ in this space and $D_{u_j^{t_2}} \in -A^{-1}(u_j^{t_2})$, the values of u_j^t can only change all at once when $w_j^t = \mathbf{0}$. We thus have from here, thanks to the second point of Condition 1, that either $w_j^t \in \mathbb{R}_-^* D_{u_j^{t_2}}$ for any $t \in [t_2, \tau]$, or $w_j^\tau = \mathbf{0}$. The latter actually implies $a_j^\tau \neq 0$ and

$$\langle D(w_j^\tau, \mathbf{0}), \frac{w_j^\tau}{a_j^\tau}\rangle \geq \|D(w_j^\tau, \mathbf{0})\|.$$

With $\mathcal{M} = A^{-1}(u_j^{t_2})$ again, the remaining cases of Equation (32) correspond to

- $w_j^{t_2} = \mathbf{0}$;
- or $D_{u_j^{t_2}} \in \mathcal{M} \cup \{\mathbf{0}\}$ and $w_j^{t_2} \neq \mathbf{0}$.

The first case directly yields w_j^τ thanks to the second point of Condition 1. The second case yields, thanks to Lemma 13 that

$$w_j^t \in \mathcal{M} \text{ for any } t \in [t_2, \tau].$$

Equation (31) then rewrites for any $t \in [t_2, \tau]$

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} \geq \|D(w_j^t, \mathbf{0})\|^2 - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 - \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2, \tag{34}$$

where $D(w_j^t, \mathbf{0})$ is constant on $[t_2, \tau]$. Solutions of the ODE $f'(t) = c^2 - f^2(t)$ with value in $(-c, c)$ are of the form $f(t) = c\tanh(c(t - t_0))$ for $t_0 \in \mathbb{R}$. A Grönwall type comparison leads to

$$\begin{aligned} &\forall t \in [t_2, \tau], \langle \mathsf{w}_j^t, D(w_j^t, 0)\rangle \geq c_j \tanh(c_j(t - t_0)) \\ &\text{where} \quad \langle \mathsf{w}_j^{t_2}, D(w_j^{t_2}, 0)\rangle = c_j \tanh(c_j(t_2 - t_0)) \\ \text{and} \quad &c_j = \sqrt{\|D(w_j^{t_2}, 0)\|^2 - \frac{2}{n}\|D(w_j^{t_2}, 0)\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2}. \end{aligned} \tag{35}$$

Since $\langle \mathsf{w}_j^{t_2}, D(w_j^{t_2},0)\rangle \geq -\|D(w_j^{t_2},0)\|\sqrt{1-\min(\alpha_{\min}^2,\alpha_0^2)}$, a simple computation using the inequality $\tanh(x)\leq -1+e^x$ yields

$$\begin{aligned}
e^{c_j(t_2-t_0)} &\geq 1 - \frac{\sqrt{1-\min(\alpha_{\min}^2,\alpha_0^2)}}{\sqrt{1-\frac{2}{n\|D(w_j^{t_2},0)\|}\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2}} \\
&\geq 1 - \frac{\sqrt{1-\min(\alpha_{\min}^2,\alpha_0^2)}}{\sqrt{1-\frac{\min(\alpha_{\min}^2,\alpha_0^2)}{2}}} \\
&\geq \frac{\min(\alpha_{\min}^2,\alpha_0^2)}{4}.
\end{aligned}$$

The second inequality comes from the choice of $\lambda_{\alpha_0}^*$, while the third one comes from the convex inequality $1-\frac{\sqrt{1-x}}{\sqrt{1-\frac{x}{2}}}\geq \frac{x}{4}$.

Using $\tanh(x)\geq 1-e^{-x}$, we now get for any $t\in[t_2,\tau]$

$$\begin{aligned}
\langle \mathsf{w}_j^t, D(w_j^t,0)\rangle &\geq c_j(1-e^{-c_j(t_2-t_0)}e^{c_j(t-t_2)}) \\
&\geq c_j\left(1-\frac{4}{\min(\alpha_{\min}^2,\alpha_0^2)}e^{-c_j(t-t_2)}\right).
\end{aligned}$$

Plugging the values of c_j, t_2 and using the choice of $\lambda_{\alpha_0}^*$, this finally leads to

$$\begin{aligned}
\langle \mathsf{w}_j^\tau, D(w_j^\tau,\mathbf{0})\rangle \geq \|D(w_j^\tau,\mathbf{0})\| &- \frac{4D_{\max}}{\min(\alpha_{\min}^2,\alpha_0^2)} e^{\frac{4D_{\max}^2}{D_{\min}^2\min(\alpha_{\min},\alpha_0)^2}} \lambda^{\frac{\|D(w_j^\tau,0)\|}{D_{\max}}\varepsilon} \lambda^{-\sqrt{\frac{2\sum_{k=1}^n\|x_k\|^2}{nD_{\max}}}\lambda^{1-2\varepsilon}\varepsilon} \\
&- \sqrt{\frac{2}{n}\|D(w_j^\tau,0)\|\sum_{k=1}^n\|x_k\|^2\lambda^{1-2\varepsilon}}.
\end{aligned}$$

This yields Theorem 1 where the hidden constant[13] c_{α_0} in the $\mathcal{O}_{\alpha_0}$ is

$$c_{\alpha_0} = \frac{4D_{\max}}{\min(\alpha_{\min}^2,\alpha_0^2)} e^{\frac{4D_{\max}^2}{D_{\min}^2\min(\alpha_{\min},\alpha_0)^2}} e^{\sqrt{\frac{2\sum_{k=1}^n\|x_k\|^2}{nD_{\max}e^2}}} + \sqrt{\frac{2}{n}D_{\max}\sum_{k=1}^n\|x_k\|^2}, \tag{36}$$

given that $t_2\leq\tau$ (otherwise, we can trivially take $c_{\alpha_0}=2D_{\max}$).

Appendix E. General Alignment Theorem

This section provides an alternative version of Theorem 1, that holds even if the function G has saddle points. In return, it requires a stronger condition on the neurons, that we explain further below.

Condition 2. *The neuron $j\in[m]$ satisfies:*

1. $\langle D(w_j^0,\mathbf{0}), \frac{w_j^0}{a_j^0}\rangle > 0$;

13. We here use the fact that $-x\ln(x)\leq\frac{1}{e}$ and $\varepsilon<\frac{1}{3}$ to deal with the $\lambda^{-\lambda}$ term.

2. for any $t \in \mathbb{R}_+$ and $\varepsilon > 0$:

$a_j^{t+\delta} D(w_j^t, \theta^{t+\delta}) \in A^{-1}(A(w_j^t))$ *for any* $\delta \in [0, \varepsilon] \implies w_j^{t+\delta} \in A^{-1}(A(w_j^t))$ *for any* $\delta \in [0, \varepsilon]$.

Theorem 3. *If Assumptions 1 and 2 hold, then the following holds almost surely for any constant $\varepsilon \in (0, \frac{1}{3})$ and initialisation scale $\lambda < \lambda_1^*$, where $\lambda_1^* > 0$ only depends[14] on the data $(x_k, y_k)_k$; with $D_{\max} \coloneqq \max_{w \in \mathbb{R}^d} \|D(w, \mathbf{0})\|$ and $\tau \coloneqq -\frac{\varepsilon \ln(\lambda)}{D_{\max}}$,*

(i) neurons' norms do not change until τ:

$$\forall t \leq \tau, \forall j \in [m], |a_j^0| \lambda^{2\varepsilon} \leq |a_j^t| \leq |a_j^0| \lambda^{-2\varepsilon}.$$

(ii) Moreover, for any neuron j satisfying Condition 2, $D(w_j^\tau, \mathbf{0})$ is an extremal vector, along which w_j^τ is aligned:

$$\langle D(w_j^\tau, \mathbf{0}), \frac{w_j^\tau}{a_j^\tau} \rangle \geq \|D(w_j^\tau, \mathbf{0})\| - \mathcal{O}\left(\lambda^{\frac{\|D(w_j^\tau, \mathbf{0})\|}{D_{\max}} \varepsilon}\right).$$

Remark 4. *The first point of Condition 2 lower bounds the quantity $\langle D(w_j^t, \mathbf{0}), \frac{w_j^t}{a_j^t} \rangle$ by 0 at initialisation and is needed for the second point of Theorem 3. Even though $\langle D(w_j^t, \mathbf{0}), \frac{w_j^t}{a_j^t} \rangle$ is increasing over time, the quantity $\|D(w_j^t, \mathbf{0})\| + \langle D(w_j^t, \mathbf{0}), \frac{w_j^t}{a_j^t} \rangle$ is not necessarily monotone, since $\|D(w_j^t, \mathbf{0})\|$ can drastically change from an activation cone to another. As a consequence, it becomes challenging to bound $\langle D(w_j^t, \mathbf{0}), \frac{w_j^t}{a_j^t} \rangle$ away from $-\|D(w_j^t, \mathbf{0})\|$ during the whole alignment phase; which would allow to lower bound the alignment rate. The $\langle D(w_j^0, \mathbf{0}), \frac{w_j^0}{a_j^0} \rangle > 0$ condition allows to do so, since it holds during the whole alignment phase by monotonicity.*

In the absence of saddle points of G, a more general condition $\langle D(w_j^0, \mathbf{0}), \frac{w_j^0}{a_j^0} \rangle > -\sqrt{1 - \alpha_0^2} \|D(w_j^t, \mathbf{0})\|$ is used. In that case, we can indeed show that $\langle D(w_j^t, \mathbf{0}), \frac{w_j^t}{a_j^t} \rangle$ is either bounded away from $-\|D(w_j^t, \mathbf{0})\|$ when entering a new activation cone, or is exactly $-\|D(w_j^t, \mathbf{0})\|$, in which case it stays aligned with $-D(w_j^t, \mathbf{0})$ for the whole phase. Such a quantisation of the possible values of $\langle D(w_j^t, \mathbf{0}), \frac{w_j^t}{a_j^t} \rangle$ when entering a new activation cone does not hold with the presence of saddle points.

Remark 5. *The second point in Condition 2 roughly states that the neuron j does not spontaneously leave an activation manifold when it can remain inside this manifold (see Example 1 in Appendix E.1 for further insights on the condition). This condition is needed to avoid degenerate situations that could happen near saddle points of G (or their corresponding activation manifolds). Lemma 13 indeed states that this point is guaranteed (at least up to a time τ, which is sufficient to prove Theorem 1) with the absence of saddle points.*

14. The exact value of λ_1^* is given by Equation (21) in Appendix D.

E.1 Understanding Condition 2

Example 1. *Consider a simplified case with a single data point $(x_k, y_k) = (1, 1)$. In the early phase, Equation (2) can then be rewritten with some positive function $g_j(t)$ as*[15]

$$\frac{\mathrm{d}w_j^t}{\mathrm{d}t} \in g_j(t)\partial\sigma(w_j^t). \tag{37}$$

In the case of ReLU activation ($\gamma = 0$) and $w_j^0 = 0$, the set of solutions of Equation (37) are described by functions of the following form, for all $t_0 \in \mathbb{R}_+ \cup \{\infty\}$,

$$w_j^t = \begin{cases} 0 \text{ for any } t \in [0, t_0] \\ \int_{t_0}^t g_j(s)\mathrm{d}s \text{ for any } t > t_0 \end{cases}.$$

In particular, only the constant solution $w_j^t \equiv 0$ satisfies the second point of Condition 2. Condition 2 indeed prevents the neuron to spontaneously leave at a finite time (given by t_0) its activation manifold (here given by $w = 0$) when it can stay within.

Although solutions satisfying the second point of Condition 2 might seem natural from a gradient flow point of view, we believe that the limits of gradient descent dynamics with arbitrarily small step size (i.e. Euler solutions) do not necessarily correspond to this kind of solutions. Non-zero step size solutions can indeed bounce from a side to another side of a stable manifold (without never being exactly located on this manifold). They could then escape this manifold if it later becomes unstable, due to the changes in the estimated function h_θ. Characterising the exact limit of gradient descent with non-differentiable activations is out of our scope and remains open for future work.

E.2 Proof of Theorem 3

(i) The first point follows the exact same lines as Appendix D.3.

(ii) Consider a neuron j that satisfies Condition 2 and $a_j^0 > 0$. The symmetric case (a_j^0) is dealt with similarly. The exact same arguments as in the proof of Theorem 1 in Appendix D.7 lead again to Equation (31) that we recall here.

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} \geq \|D(w_j^t, \mathbf{0})\|^2 - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 - \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2.$$

We define again $u_j^t = A(w_j^t)$ and $D_{u_j^t} \coloneqq D(w_j^t, \mathbf{0}) \in \mathfrak{D}_{u_j^t}$ in the following. If $D_{u_j^t} \notin A^{-1}(u_j^t) \cup \{\mathbf{0}\}$, the definition of $\alpha_{\min,+}$ (which is positive thanks to Lemma 9), given by Equation (22), implies that

$$\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle \leq \|D(w_j^t, \mathbf{0})\|\sqrt{1 - \alpha_{\min,+}^2}.$$

Equation (31) implies that as long as $\|D(w_j^t, \mathbf{0})\|^2 - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 \geq \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2$, the quantity $\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle$ is increasing. In particular, the choice of λ makes it increasing

15. To be rigorous $g_j(t)$ should also depend on $(w_j^s)_{s<t}$, but this does not change the point.

as long as $D_{u_j^t} \notin A^{-1}(u_j^t)$ and $\langle \mathsf{w}_j^t, D_{u_j^t} \rangle > 0$. Since $\langle \mathsf{w}_j^0, D_{u_j^0} \rangle > 0$ by Condition 2, we have again by continuous induction that as long as $D_{u_j^t} \notin A^{-1}(u_j^t)$:

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} \geq \frac{D_{\min}^2}{2}\alpha_{\min}^2. \tag{38}$$

This then implies that after a time t_2, $D_{u_j^{t_2}} \in A^{-1}(u_j^{t_2})$ with

$$t_2 \leq \frac{2D_{\max}}{D_{\min}^2 \alpha_{\min}^2}.$$

Again, we can observe that either $\lambda^\varepsilon = \Omega(1)$ or $t_2 \leq \tau$ and we focus only on the latter case. From there, Lemma 11, proven in Appendix D, implies that for any $t \in [t_2, \tau]$, $D(w_j^{t_2}, \theta^t) \in A^{-1}(u_j^{t_2})$. The second point of Condition 2 then implies that $w_j^t \in A^{-1}(u_j^{t_2})$ for any $t \in [t_2, \tau]$. Equation (31) now rewrites

$$\frac{\mathrm{d}\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle}{\mathrm{d}t} \geq \|D(w_j^t, \mathbf{0})\|^2 - \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle^2 - \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2,$$

where $D(w_j^t, \mathbf{0})$ is constant on $[t_2, \tau]$. Again, comparing with solutions of the ODE $f'(t) = c^2 - f^2(t)$, we get

$$\forall t \in [t_2, \tau], \langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle \geq c_j \tanh(c_j(t - t_2))$$
$$\text{where} \quad c_j = \sqrt{\|D(w_j^t, \mathbf{0})\|^2 - \frac{2}{n}\|D(w_j^t, \mathbf{0})\|\lambda^{2-4\varepsilon}\sum_{k=1}^n \|x_k\|^2}.$$

Since $\tanh(x) \geq 1 - e^{-x}$, the previous equation rewrites for any $t \in [t_2, \tau]$

$$\langle \mathsf{w}_j^t, D(w_j^t, \mathbf{0})\rangle \geq c_j\left(1 - e^{-c_j(t-t_2)}\right).$$

Using the value of c_j and bound on t_2,

$$\begin{aligned}\langle \mathsf{w}_j^\tau, D(w_j^\tau, \mathbf{0})\rangle &\geq c_j\left(1 - e^{c_j t_2}\lambda^{\frac{c_j}{D_{\max}}\varepsilon}\right) \\ &\geq \|D(w_j^\tau, \mathbf{0})\| - D_{\max}e^{\frac{2D_{\max}^2}{D_{\min}^2\alpha_{\min}^2}}\lambda^{\frac{\|D(w_j^\tau,0)\|}{D_{\max}}\varepsilon}\lambda^{-\sqrt{\frac{2\sum_{k=1}^n\|x_k\|^2}{nD_{\max}}}\lambda^{1-2\varepsilon}\varepsilon} \\ &\quad - \sqrt{\frac{2}{n}\|D(w_j^\tau, 0)\|\sum_{k=1}^n\|x_k\|^2}\lambda^{1-2\varepsilon}.\end{aligned}$$

This leads to the second part of Theorem 3 where the hidden constant c in the $\mathcal{O}$ is

$$c = D_{\max}e^{\frac{2D_{\max}^2}{D_{\min}^2\alpha_{\min}^2}}e^{\sqrt{\frac{2\sum_{k=1}^n\|x_k\|^2}{nD_{\max}e^2}}} + \sqrt{\frac{2}{n}D_{\max}\sum_{k=1}^n\|x_k\|^2}, \tag{39}$$

given that $t_2 \leq \tau$ (otherwise, we take $c = 2D_{\max}$).

E.3 Absence of Saddle Points for $d \leq 2$.

Footnote 4 states that saddles points of G only exist on sectors of dimension at most $d-2$. A rigorous statement of this affirmation is given by Lemma 15 below. For that we need an assumption on the data generation, that is slightly weaker than Assumption 2.

Assumption 7. *The data (x_k, y_k) are drawn independently at random, following a distribution such that for any $k, k' \in [n]$, the random variable $y_k \langle \frac{x_{k'}}{\|x_{k'}\|}, x_k \rangle$ admits a density in $\mathbb{R}$.*

Lemma 15. *If $d \leq 2$ and Assumption 7 holds, then for any stationary point $w^* \in \mathbb{S}_d$ of the function G defined in Equation (7), w^* is a local minimum of G on $\mathbb{S}_d$.*

Proof. If $d = 1$, then $\mathbb{S}_d$ is discrete and so any point is a local minimum.

Consider now $d = 2$ and a stationary point w^* of G. First assume that $\langle w^*, x_k \rangle \neq 0$ for any $k \in [n]$ such that $y_k x_k \neq 0$. Then G is by definition linear in a neighbourhood of w^*, so that w^* is not a saddle point of G.

Now assume instead that $\mathrm{Span}(\{x_k \mid \langle w^*, x_k \rangle = 0\})$ has dimension 1—it cannot be of dimension 2 since $w^* \neq \mathbf{0}$. Even if it means reordering the data, we can assume that $x_1 \neq 0$, $\langle x_1, w^* \rangle = 0$. Moreover, Assumption 7 implies that almost surely, for any $k \geq 2$, (x_1, x_k) are unaligned and so $\langle w^*, x_k \rangle \neq 0$. Then note that $\mathbb{R}^d = \mathrm{Span}(x_1) \oplus^\perp \mathrm{Span}(w^*)$. In consequence, we can consider the following neighbourhood of w^* in $\mathbb{S}_d$ for any δ:

$$N(\delta) = \left\{ \cos(\theta) w^* + \frac{\sin(\theta)}{\|x_1\|} x_1 \mid |\theta| < \delta \right\}. \tag{40}$$

At proximity of w^* (i.e., for small enough θ),

$$\begin{aligned} G\left(\cos(\theta) w^* + \frac{\sin(\theta)}{\|x_1\|} x_1\right) &= G(w^*) + \sin(\theta)_+ \frac{y_1 \|x_1\|}{n} + \frac{1}{n} \sum_{k \geq 2} y_k \sin(\theta) \langle \frac{x_1}{\|x_1\|}, x_k \rangle \qquad (41) \\ &\quad + \frac{1 - \cos(\theta)}{n} \sum_{k \geq 2} y_k \langle w^*, x_k \rangle \\ &= G(w^*) + g_1(\theta) + \mathcal{O}\left(\theta^2\right), \qquad (42) \end{aligned}$$

where $g_1(t) = \frac{y_1 \|x_1\|}{n}(t)_+ + \frac{t}{n} \sum_{k \geq 2} y_k \langle \frac{x_1}{\|x_1\|}, x_k \rangle$. Note that g_1 is a continuous, piecewise affine function on $\mathbb{R}$. Moreover, both its left and right slopes at 0 are almost surely non-zero, thanks to Assumption 7. As a consequence, the piecewise affine terms will dominate in Equation (42), and w^* is necessarily a local minimum—this comes from the fact that a piecewise affine function in $\mathbb{R}$ does not admit any saddle point. □

Appendix F. Proof of Theorem 2

We prove in this section a more general version of Theorem 2, given by Theorem 4 below.

Theorem 4. *If Assumptions 4, 5 and 6 hold, then there exists some constant $\tilde{\lambda} = \Theta(1)$ such that for any $\lambda < \tilde{\lambda}$ and $m \in \mathbb{N}$, the parameters θ^t converge to some θ^∞_λ such that*

$$h_{\theta^\infty_\lambda}(x) = x^\top \beta^* \text{ for any } x \in \mathcal{C},$$

where $\mathcal{C} = \{tz \mid t \in \mathbb{R}_+, z \in \mathrm{Conv}(\{x_k \mid k \in [n]\})\}$ *is the cone generated by the convex hull of the data points. Moreover*

$$\forall x \in \mathbb{R}^2, \lim_{\lambda \to 0^+} h_{\theta^\infty_\lambda}(x) = (x^\top \beta^*)_+.$$

In the following, we prove Lemmas 4, 5, 6 and 7, but in more general versions, as follows:

- Lemma 4 only require Assumption 4 (instead of Assumption 3).
- Lemmas 5 and 6 only require Assumptions 4 and 5.
- Lemma 7 only require Assumptions 4, 5 and 6.

F.1 Phase 1

Proof of Lemma 4. Note in this proof $\mathsf{w}_i^t = \frac{w_i^t}{a_i^t}$. A crucial observation is that the only non-zero extremal vector is D^*, which corresponds to a local maximum of G. Assumption 4 actually implies Lemmas 2, 9, 10 and 11 (and the absence of saddle points), that are sufficient (instead of Assumption 2), to apply[16] Theorem 1. Showing these implications is direct once we observe D^* is the only extremal vector and is omitted for the sake of conciseness.

1) The only non-zero extremal vector is D^*, which corresponds to a local maximum of G. Also notice that for every $i \in \mathcal{I}$, Assumption 4 implies that $\langle D(w_i^0, 0), \mathsf{w}_i^0 \rangle > 0$. Moreover, we can show that $\langle D(w, \theta^t), w \rangle \geq 0$ for any $t \leq \tau$ and $w \in \mathbb{R}^d$, so that a_i^t is non-decreasing during the early phase for $i \in \mathcal{I}$. This yields $w_i^t \neq \mathbf{0}$ for $i \in \mathcal{I}$ and $t \leq \tau$. As a consequence, any neuron $i \in \mathcal{I}$ satisfies Condition 1. Theorem 1 can then be applied and implies the neuron i is aligned with D^* at time τ i.e.,

$$\langle D^*, \mathsf{w}_i^\tau \rangle \geq \|D^*\| - c\lambda^\varepsilon.$$

Moreover, $a_i^0 \leq a_i^\tau \leq a_i^0 \lambda^{-2\varepsilon}$ follows by Theorem 1 and noticing that a_i^t is non-decreasing on $[0, \tau]$ for any $i \in \mathcal{I}$.

2) For the same reason as in 1), a_i^t is non-decreasing, and non-positive thanks to Lemma 1.

3) By definition of $\mathcal{I}$ and $\mathcal{N}$, note that a.s. for $i \notin \mathcal{I} \cup \mathcal{N}$, $\langle w_i^0, x_k \rangle < 0$ for any $k \in [n]$. As a consequence, any neuron $i \notin \mathcal{I} \cup \mathcal{N}$ can be ignored, since its gradient is null during the whole training: $\mathfrak{D}(w_i^t, \theta^t) = \{\mathbf{0}\}$ for any $t \geq 0$. □

F.2 Phase 2a

Proof of Lemma 5. Consider some $\varepsilon_2 > 0$ and define

$$\tau_2 \coloneqq \inf \left\{ t \geq \tau \mid \sum_{i \in \mathcal{I}} (a_i^t)^2 \geq \varepsilon_2 \text{ or } \sum_{i \in \mathcal{N}} (a_i^t)^2 \geq 2\lambda^2 \right\}.$$

16. Otherwise, we could just state Theorem 2 for almost all the datasets in the considered open set.

First, for any $t \in [\tau, \tau_2]$, $-\mathcal{O}\left(\lambda^2 y_k\right) \leq h_{\theta^t}(x_k) \leq \mathcal{O}\left(\varepsilon_2 y_k\right)$ for any k. Thanks to Assumption 4 and our choice of λ, this implies the following equality for any $t \in [\tau, \tau_2]$

$$D^t = \|D^*\| + \mathcal{O}\left(\varepsilon_2\right)$$

As a consequence, for a small enough $\varepsilon_2^* = \Theta(1)$ such that $\varepsilon_2 \leq \varepsilon_2^*$ and positive constants $\delta_k = \langle \frac{D^*}{\|D^*\|}, x_k \rangle - \Theta(\varepsilon_2)$, this yields

$$D^t \in \left\{ u \in \mathbb{R}^d \mid \forall k \in [n], \langle \frac{u}{\|u\|}, x_k \rangle \geq \delta_k \right\}.$$

From there, Equation (30) implies that for any $t \in [\tau, \tau_2]$ and as long as $\langle \mathsf{w}_i^t, x_k \rangle > 0$ for any k:

$$\frac{\mathrm{d}\langle \mathsf{w}_i^t, x_k \rangle}{\mathrm{d}t} \geq \|D^t\| \left(\delta_k - \langle \mathsf{w}_i^t, x_k \rangle\right).$$

Moreover, $\langle \mathsf{w}_i^\tau, x_k \rangle \geq \delta_k$ thanks to Lemmas 4 and 16 for a small enough $\frac{\lambda^{\frac{\varepsilon}{2}}}{\varepsilon_2}$. This last ODE then yields that $\langle \mathsf{w}_j^t, x_k \rangle \geq \delta_k$ for any $k \in [n]$ and $t \in [\tau, \tau_2]$, which is the second point of Lemma 5.

Also we can choose $\varepsilon_2^* = \Theta(1)$ and $\tilde{\lambda} = \Theta(1)$ small enough so $|h_{\theta^t}(x_k)| < y_k$ for any $t \in [\tau, \tau_2]$. This then implies that a_i^t is non-decreasing for any $i \in [m]$ on $[\tau, \tau_2]$. In particular, this is the case for $i \in \mathcal{N}$, which implies the fourth point of Lemma 5.

Moreover, we can bound the growth of a_i^t for $i \in \mathcal{I}$. In particular, we have for $i \in \mathcal{I}$ and $t \in [\tau, \tau_2]$,

$$\frac{\mathrm{d}a_i^t}{\mathrm{d}t} = a_i^t \langle \mathsf{w}_i^t, D^t \rangle \geq \frac{a_i^t}{n} \sum_{k=1}^n \delta_k (1 - \mathcal{O}\left(\varepsilon_2\right)) y_k. \tag{43}$$

This implies an exponential lower bound on the growth of a_i^t, and so we have $\tau_2 < +\infty$. Since the second condition in the definition of τ_2 does not break at τ_2 (thanks to the fourth point already shown above), the first condition necessarily breaks at τ_2. By continuity, this implies the first point of Lemma 5: $\sum_{i \in \mathcal{I}} (a_i^{\tau_2})^2 = \varepsilon_2$.

It now remains to prove the third point of Lemma 5. We can also upper bound the growth of a_i^t on $[\tau, \tau_2]$:

$$\begin{aligned}
\frac{\mathrm{d}a_i^t}{\mathrm{d}t} &\leq \frac{a_i^t}{n} \sum_{k=1}^n (1 + \mathcal{O}\left(\lambda^2\right)) y_k \langle \mathsf{w}_i^t, x_k \rangle \\
&= a_i^t (1 + \mathcal{O}\left(\lambda^2\right)) \langle \mathsf{w}_i^t, D^* \rangle \\
&\leq a_i^t (1 + \mathcal{O}\left(\lambda^2\right)) \|D^*\|.
\end{aligned}$$

A Grönwall comparison argument then allows to write, with ε given by Lemma 4:

$$\begin{aligned}
\sum_{i \in \mathcal{I}} (a_i^{\tau_2})^2 &\leq e^{2(1+\mathcal{O}\left(\lambda^2\right))\|D^*\|(\tau_2 - \tau)} \sum_{i \in \mathcal{I}} (a_i^\tau)^2 \\
&\leq e^{2(1+\mathcal{O}\left(\lambda^2\right))\|D^*\|(\tau_2 - \tau)} \lambda^{2-4\varepsilon}.
\end{aligned}$$

Since the left term is exactly ε_2, this gives the following bound on τ_2:

$$\begin{aligned}\tau_2 - \tau &\geq -\frac{1-2\varepsilon}{\|D^*\|}\ln(\lambda) - \mathcal{O}\left(\ln(\frac{1}{\varepsilon_2}) + \lambda^2\right)\\ &\geq -\frac{1-2\varepsilon}{\|D^*\|}\ln(\lambda) - \mathcal{O}\left(\ln(\frac{1}{\varepsilon_2})\right). \end{aligned} \tag{44}$$

Moreover, a similar bound to Equation (43) yields for any $i \in \mathcal{I}$ and $t \in [\tau, \tau_2]$:

$$\begin{aligned}\langle \mathsf{w}_i^t, D^t\rangle &\geq \frac{1}{n}\sum_{k=1}^{n}(1-\mathcal{O}(\varepsilon_2))\delta_k y_k\\ &= (1-\mathcal{O}(\varepsilon_2))\langle \frac{D^*}{\|D^*\|}, D^*\rangle\\ &= (1-\mathcal{O}(\varepsilon_2))\|D^*\|.\end{aligned}$$

Let $i, j \in \mathcal{I}$, we can now study the evolution of $\langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle$ for $t \in [\tau, \tau_2]$:

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle}{\mathrm{d}t} &= \langle D^t, \mathsf{w}_i^t + \mathsf{w}_j^t\rangle\left(1 - \langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle\right)\\ &\geq 2(1-\mathcal{O}(\varepsilon_2))\|D^*\|\left(1 - \langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle\right).\end{aligned}$$

Again, a Grönwall comparison argument implies, using the bound on $\tau_2 - \tau$ in Equation (44),

$$\begin{aligned}1 - \langle \mathsf{w}_i^{\tau_2}, \mathsf{w}_j^{\tau_2}\rangle &\leq e^{-2(1-\mathcal{O}(\varepsilon_2))\|D^*\|(\tau_2-\tau)}\left(1 - \langle \mathsf{w}_i^{\tau}, \mathsf{w}_j^{\tau}\rangle\right)\\ &\leq e^{-(1-\mathcal{O}(\varepsilon_2))\left(-(2-4\varepsilon)\ln(\lambda) - \mathcal{O}\left(\ln(\frac{1}{\varepsilon_2})\right)\right)}\left(1 - \langle \mathsf{w}_i^{\tau}, \mathsf{w}_j^{\tau}\rangle\right)\\ &\leq \mathcal{O}\left(\varepsilon_2^{-1+\mathcal{O}(\varepsilon_2)}\lambda^{(1-\mathcal{O}(\varepsilon_2))(2-4\varepsilon)}\right)\left(1 - \langle \mathsf{w}_i^{\tau}, \mathsf{w}_j^{\tau}\rangle\right).\end{aligned} \tag{45}$$

Moreover, we can bound $\left(1 - \langle \mathsf{w}_i^{\tau}, \mathsf{w}_j^{\tau}\rangle\right)$, thanks to Lemma 4. We can indeed write

$$\begin{gathered}\mathsf{w}_i^{\tau} = \langle \mathsf{w}_i^{\tau}, \frac{D^*}{\|D^*\|}\rangle\frac{D^*}{\|D^*\|} + v_i^{\tau}\\ \text{where } \langle v_i^{\tau}, D^*\rangle = 0 \text{ and } \|v_i^{\tau}\|^2 = \mathcal{O}(\lambda^{\varepsilon}).\end{gathered}$$

This then implies

$$\begin{aligned}\langle \mathsf{w}_i^{\tau}, \mathsf{w}_j^{\tau}\rangle &= \langle \mathsf{w}_i^{\tau}, \frac{D^*}{\|D^*\|}\rangle\langle \mathsf{w}_j^{\tau}, \frac{D^*}{\|D^*\|}\rangle + \langle v_i^{\tau}, v_j^{\tau}\rangle\\ &\geq 1 - \mathcal{O}(\lambda^{\varepsilon}).\end{aligned}$$

Equation (45) then rewrites

$$\begin{aligned}1 - \langle \mathsf{w}_i^{\tau_2}, \mathsf{w}_j^{\tau_2}\rangle &\leq \mathcal{O}\left(\varepsilon_2^{-1+\mathcal{O}(\varepsilon_2)}\lambda^{(1-\mathcal{O}(\varepsilon_2))(2-4\varepsilon)+\varepsilon}\right)\\ &\leq \mathcal{O}\left(\varepsilon_2^{-1}\lambda^{(1-\mathcal{O}(\varepsilon_2))(2-4\varepsilon)+\varepsilon}\right).\end{aligned}$$

The third point of Lemma 5 then follows for a small enough choice of $\varepsilon_2^* = \Theta(1)$ (that can depend on ε) and $\varepsilon_2 = \Omega(\lambda^{\varepsilon})$. □

Lemma 16. *Consider ε, τ defined in Lemma 4. For any $\delta_k = \langle \frac{D^*}{\|D^*\|}, x_k \rangle - \Theta(\varepsilon_2)$, there exists a large enough constant c' such that if Assumption 4 holds and $\varepsilon_2 \geq c' \lambda^{\frac{\varepsilon}{2}}$, then for all $i \in \mathcal{I}$:*

$$\forall k \in [n], \langle \mathsf{w}_i^\tau, x_k \rangle \geq \delta_k,$$

where δ_k is defined in the proof of Lemma 5.

Proof. For any $i \in \mathcal{I}$, we use the same decomposition as in the proof of Lemma 5,

$$\mathsf{w}_i^\tau = \langle \mathsf{w}_i^\tau, \frac{D^*}{\|D^*\|} \rangle \frac{D^*}{\|D^*\|} + v_i^\tau$$
$$\text{where } \langle v_i^\tau, D^* \rangle = 0 \text{ and } \|v_i^\tau\|^2 = \mathcal{O}\left(\lambda^\varepsilon\right).$$

This decomposition yields thanks to Lemma 4, for some data dependent constant $c > 0$

$$\begin{aligned}
\langle \mathsf{w}_i^\tau, x_k \rangle &\geq \left(1 - \mathcal{O}\left(\lambda^\varepsilon\right)\right) \langle \frac{D^*}{\|D^*\|}, x_k \rangle - \|v_i^\tau\| \|x_k\| \\
&\geq \left(1 - \frac{c}{\|D^*\|} \lambda^\varepsilon\right) \langle \frac{D^*}{\|D^*\|}, x_k \rangle - \mathcal{O}\left(\lambda^{\frac{\varepsilon}{2}}\right) \\
&\geq \langle \frac{D^*}{\|D^*\|}, x_k \rangle - \mathcal{O}\left(\lambda^{\frac{\varepsilon}{2}}\right).
\end{aligned}$$

We thus have for a large enough c' with $\varepsilon_2 \geq c' \lambda^{\frac{\varepsilon}{2}}$,

$$\langle \mathsf{w}_i^\tau, x_k \rangle \geq \delta_k.$$

□

F.3 Phase 2b

Define in the following

$$\beta \coloneqq \min_{k \in \mathcal{K}} \frac{y_k \|x_k^2\|}{n} - \frac{1}{n} \sqrt{\sum_{k'=1}^n y_{k'}^2} \sqrt{\sum_{k' \neq k} \langle x_k', x_k \rangle^2}, \tag{46}$$

(47)

Assumption 5 implies that β is positive.

Proof of Lemma 6. For this phase, define for some $\delta' = \Theta(1)$

$$\begin{aligned}
\tau_3 \coloneqq \inf \big\{ t \geq \tau_2 \mid \left\| \beta^* - \sum_{i \in I \cup N} a_i^t w_i^t \right\| \leq \varepsilon_3 \text{ or } \exists i, j \in \mathcal{I}, \langle \mathsf{w}_i^t, \mathsf{w}_j^t \rangle \leq 1 - \lambda^\varepsilon \\
\text{or } \exists i \in \mathcal{I}, \exists k \in [n], \langle \mathsf{w}_i^t, x_k \rangle \leq \delta' \|x_k\| \text{ or } \sum_{i \in \mathcal{N}} (a_i^t)^2 \geq \lambda^{2(1-\varepsilon)} \big\}.
\end{aligned}$$

Thanks to Lemmas 18 and 19 proven below, we have for any $t \in [\tau_2, \tau_3]$:

$$\forall i \in \mathcal{I}, \forall k \in [n], \langle \mathsf{w}_i^t, x_k \rangle > \delta' \|x_k\|$$
$$\text{and } \sum_{i \in \mathcal{I}} (a_i^t)^2 \geq \varepsilon_2.$$

We note in the following $\beta^t_{\mathcal{I}} = \sum_{i\in\mathcal{I}} a^t_i w^t_i$, $H = \frac{1}{n}X^\top X$ and for any vector $w \in \mathbb{R}^d$, $\|w\|^2_H = w^\top H w$. The distance between β^* and $\beta^t_{\mathcal{I}}$ then decreases as

$$\frac{\mathrm{d}\|\beta^* - \beta^t_{\mathcal{I}}\|^2_H}{\mathrm{d}t} = 2(\beta^t_{\mathcal{I}} - \beta^*)^\top H \frac{\mathrm{d}\beta^t_{\mathcal{I}}}{\mathrm{d}t} \tag{48}$$

$$= 2(\beta^t_{\mathcal{I}} - \beta^*)^\top H \left(\sum_{i\in\mathcal{I}} (a^t_i)^2 I_d + w^t_i w^{t\ \top}_i\right) D^t. \tag{49}$$

Equation (49) comes from the fact that neurons in $\mathcal{I}$ are activated along all data points. Also, note that for any $t \in [\tau_2, \tau_3]$

$$\begin{aligned}
D^t &= -\frac{1}{n}\sum_{k=1}^n (h^{\theta^t}(x_k) - y_k)x_k \\
&= -\frac{1}{n}\sum_{k=1}^n \left((\langle \beta^t_{\mathcal{I}}, x_k\rangle - y_k)x_k + \sum_{i\in\mathcal{N}} a^t_i \langle w^t_i, x_k\rangle_+ x_k\right) \\
&= -\frac{1}{n}X^\top(X\beta^t_{\mathcal{I}} - \mathbf{y}) + \mathcal{O}\left(\lambda^{2-2\varepsilon}\right) \\
&= -H(\beta^t_{\mathcal{I}} - \beta^*) + \mathcal{O}\left(\lambda^{2-2\varepsilon}\right),
\end{aligned}$$

For the sake of clarity, we denote in the following $\sigma_{\min}$ and $\sigma_{\max}$ respectively for the smallest and largest eigenvalue of H. Equation (49) now rewrites with Lemma 17

$$\begin{aligned}
\frac{\mathrm{d}\|\beta^* - \beta^t_{\mathcal{I}}\|^2_H}{\mathrm{d}t} &= -2(\beta^t_{\mathcal{I}} - \beta^*)^\top H \left(\sum_{i\in\mathcal{I}} (a^t_i)^2 I_d + w^t_i w^{t\ \top}_i\right)\left(H(\beta^t_{\mathcal{I}} - \beta^*) - \mathcal{O}\left(\lambda^{2-2\varepsilon}\right)\right) \\
&\le -2\varepsilon_2(\beta^t_{\mathcal{I}} - \beta^*)^\top H^2(\beta^t_{\mathcal{I}} - \beta^*) + \mathcal{O}\left(\|\beta^* - \beta^t_{\mathcal{I}}\|_H \lambda^{2-2\varepsilon}\right) \\
&\le -2\varepsilon_2\sigma_{\min}\|\beta^t_{\mathcal{I}} - \beta^*\|^2_H + \mathcal{O}\left(\|\beta^* - \beta^t_{\mathcal{I}}\|_H \lambda^{2-2\varepsilon}\right).
\end{aligned}$$

This finally yields the comparison for any $t \in [\tau_2, \tau_3]$

$$\frac{\mathrm{d}\|\beta^* - \beta^t_{\mathcal{I}}\|_H}{\mathrm{d}t} \le -\varepsilon_2\sigma_{\min}\|\beta^t_{\mathcal{I}} - \beta^*\|_H + \mathcal{O}\left(\lambda^{2-2\varepsilon}\right).$$

A Grönwall comparison argument gives for $t \in [\tau_2, \tau_3]$

$$\|\beta^* - \beta^t_{\mathcal{I}}\|_H \le \|\beta^{\tau_2}_{\mathcal{I}} - \beta^*\|_H e^{-\varepsilon_2\sigma_{\min}(t-\tau_2)} + \mathcal{O}\left(\frac{\lambda^{2-2\varepsilon}}{\varepsilon_2}\right).$$

As $\varepsilon_2 = \Omega(\lambda^{\varepsilon/2})$, using the comparison between $\|\cdot\|_2$ and $\|\cdot\|_H$ norms,

$$\begin{aligned}
\|\beta^* - \beta^t_{\mathcal{I}}\|_2 &\le \frac{1}{\sqrt{\sigma_{\min}}}\|\beta^* - \beta^t_{\mathcal{I}}\|_H \\
&\le \frac{1}{\sqrt{\sigma_{\min}}}\|\beta^* - \beta^{\tau_2}_{\mathcal{I}}\|_H e^{-\varepsilon_2\sigma_{\min}(t-\tau_2)} + \mathcal{O}\left(\lambda^{2-\frac{5}{2}\varepsilon}\right) \\
&\le \left(\sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}}\|\beta^*\| + \mathcal{O}\left(\varepsilon_2\right)\right) e^{-\varepsilon_2\sigma_{\min}(t-\tau_2)} + \mathcal{O}\left(\lambda^{2-\frac{5}{2}\varepsilon}\right).
\end{aligned}$$

After some time, the first condition in the definition necessarily breaks. This allows to bound τ_3, thanks to our choice of λ and ε_3, which finally yields the bound

$$\tau_3 - \tau_2 \leq \frac{1}{\sigma_{\min}\varepsilon_2} \ln \left(\frac{\kappa\|\beta^*\| + \mathcal{O}(\varepsilon_2)}{\varepsilon_3 - \mathcal{O}\left(\lambda^{2-\frac{5}{2}\varepsilon}\right)} \right)_+, \tag{50}$$

where $\kappa = \sqrt{\frac{\sigma_{\max}}{\sigma_{\min}}}$ is the condition number of the matrix X.

We already know that the third condition in the definition of τ_3 does not break at τ_3, thanks to Lemma 18. The remaining of the proof shows that neither the second nor fourth condition can break at τ_3, given the *small* amount of time given between τ_2 and τ_3.

For the second condition, by following the ODE computed in the proof of Lemma 5 for any $i, j \in I$,

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle}{\mathrm{d}t} &\geq -2\|D^t\|(1 - \langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle) \\ &\geq -2\overline{D}(1 - \langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle),\end{aligned}$$

where $\overline{D} = \sqrt{\frac{1}{n}\sum_{k=1}^n y_k^2}\sqrt{\frac{1}{n}\sum_{k=1}^n x_k^2} + \mathcal{O}(\lambda)$, thanks to the bound on $\|D^t\|$ given in the proof of Lemma 18. From there, a Grönwall comparison directly yields for any $t \in [\tau_2, \tau_3]$

$$\begin{aligned}\langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle &\geq 1 - \left(1 - \langle \mathsf{w}_i^{\tau_2}, \mathsf{w}_j^{\tau_2}\rangle\right) e^{2\overline{D}(t-\tau_2)} \\ &\geq 1 - \mathcal{O}\left(\lambda^{1-\varepsilon}\right) \max\left(1, \left(\frac{\kappa\beta^* + \mathcal{O}(\varepsilon_2)}{\varepsilon_3 - \mathcal{O}\left(\lambda^{2-\frac{5}{2}\varepsilon}\right)}\right)^{\frac{2\overline{D}}{\sigma_{\min}\varepsilon_2}}\right),\end{aligned}$$

where we used in the second inequality the bound on $\tau_3 - \tau_2$ given by Equation (50) and the bound on $\langle \mathsf{w}_i^{\tau_2}, \mathsf{w}_j^{\tau_2}\rangle$ given by Lemma 5. Now, thanks to the choice of λ, ε_2 and ε_3, for a small enough $c_3 > 0$ depending only on the data

$$\max\left(1, \left(\frac{\kappa\beta^* + \mathcal{O}(\varepsilon_2)}{\varepsilon_3 - \mathcal{O}\left(\lambda^{2-\frac{5}{2}\varepsilon}\right)}\right)^{\frac{2\overline{D}}{\sigma_{\min}\varepsilon_2}}\right) = \mathcal{O}\left(\lambda^{-\varepsilon}\right)$$

and so

$$\langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle \geq 1 - \mathcal{O}\left(\lambda^{1-2\varepsilon}\right).$$

Thus, the second condition in the definition of τ_3 does not break for a small enough $\tilde{\lambda}$ since $\varepsilon < \frac{1}{3}$.

We now show that the fourth condition does not break either. For any $i \in \mathcal{N}$, $a_i^t \leq 0$ and we can bound the (absolute) growth of a_i^t as

$$\begin{aligned}-\frac{\mathrm{d}a_i^t}{\mathrm{d}t} &\leq -a_i^t\|D_i^t\| \\ &\leq -\overline{D}a_i^t,\end{aligned}$$

as the bound on $\|D^t\|$ actually also holds for any D_i^t. From there, using Lemma 5, for any $t \in [\tau_2, \tau_3]$:

$$-a_i^t \leq -a_i^{\tau_2} e^{\overline{D}(t-\tau_2)}.$$

Here again, this yields[17]

$$-a_i^t = -a_i^0 \mathcal{O}\left(\lambda^{-\frac{\varepsilon}{2}}\right)$$

This gives the third point of Lemma 6 and also shows that the fourth condition in the definition of τ_3 does not break for a small enough $\tilde{\lambda} = \Theta(1)$. Necessarily, the first condition in the definition in τ_3 breaks at τ_3, which ends the proof of Lemma 6. □

Lemma 17. *If $\lambda < \tilde{\lambda}$ for a small enough $\tilde{\lambda} = \Theta(1)$, for any $t \in [\tau_2, \tau_3]$*

$$\sum_{i \in \mathcal{I}} (a_i^t)^2 = \mathcal{O}(1).$$

Proof. Define

$$\alpha \coloneqq \inf_{\substack{w \in \mathbb{S}_d \\ \forall k \in [n], \langle w, x_k \rangle \geq 0}} \frac{1}{n} \sum_{k=1}^n \langle w, x_k \rangle. \tag{51}$$

First note that α is positive. It is defined as the infimum of a continuous function on a non-empty (thanks to Assumption 4) compact set, thus its minimum is reached for some $w^* \in \mathbb{S}_d$ such that $\langle w^*, x_k \rangle \geq 0$ for any k. Since the vectors $(x_k)_k$ span the whole space $\mathbb{R}^d$, the scalar product $\langle w^*, x_k \rangle$ is zero for all k if and only if $w^* = 0$. This necessarily implies that $\alpha > 0$.

Now, the decreasing of the loss over time yields for any $t \geq 0$:

$$\begin{aligned}\sum_{k=1}^n (h_{\theta^t}(x_k) - y_k)^2 &\leq \sum_{k=1}^n (h_{\theta^0}(x_k) - y_k)^2 \\ &= \sum_{k=1}^n y_k^2 + \mathcal{O}\left(\lambda^2\right).\end{aligned}$$

The triangle inequality then implies

$$\left(\sqrt{\sum_{k=1}^n h_{\theta^t}(x_k)^2} - \sqrt{\sum_{k=1}^n y_k^2}\right)^2 = \mathcal{O}(1).$$

17. We indeed just showed in the previous computations that $e^{2\overline{D}(t-\tau_2)} = \mathcal{O}\left(\lambda^{-\varepsilon}\right)$ for $t \in [\tau_2, \tau_3]$.

Moreover, note that by comparison between 1 and 2 norms:

$$\begin{aligned}\sqrt{n\sum_{k=1}^{n} h_{\theta^t}(x_k)^2} &\geq \sum_{k=1}^{n} h_{\theta^t}(x_k)\\ &= \sum_{k=1}^{n}\sum_{i\in\mathcal{I}} a_i^t\langle w_i^t, x_k\rangle + \sum_{k=1}^{n}\sum_{i\in N} a_i^t\langle w_i^t, x_k\rangle_+\\ &\geq n\alpha\sum_{i\in\mathcal{I}} a_i^t\|w_i^t\| - \mathcal{O}\left(\lambda^{2(1-\varepsilon)}\right).\end{aligned}$$

The last inequality comes from the fact that for $t\in[\tau_2,\tau_3]$, for any $i\in\mathcal{I}$ and $k\in[n]$, $\langle w_i^t, x_k\rangle\geq 0$; which allows to bound using α defined above.

Moreover, using the balancedness property (Lemma 1), $\|w_i^t\|\geq a_i^t - a_i^0$, which gives

$$\begin{aligned}\sum_{i\in\mathcal{I}} a_i^t\|w_i^t\| &\geq \sum_{i\in\mathcal{I}}(a_i^t)^2 - \sum_{i\in\mathcal{I}} a_i^t a_i^0\\ &\geq \sum_{i\in\mathcal{I}}(a_i^t)^2 - \sqrt{\sum_{i\in\mathcal{I}}(a_i^0)^2}\sqrt{\sum_{i\in\mathcal{I}}(a_i^t)^2}\\ &\geq \sum_{i\in\mathcal{I}}(a_i^t)^2 - \lambda\sqrt{\sum_{i\in\mathcal{I}}(a_i^t)^2}.\end{aligned}$$

Wrapping up the different inequalities, we finally have for $t\in[\tau_2,\tau_3]$:

$$\sum_{i\in\mathcal{I}}(a_i^t)^2 - \lambda\sqrt{\sum_{i\in\mathcal{I}}(a_i^t)^2} = \mathcal{O}(1).$$

This then implies that $\sum_{i\in\mathcal{I}}(a_i^t)^2 = \mathcal{O}(1)$. □

Lemma 18. *For small enough $\tilde{\lambda}=\Theta(1)$ and $\delta'=\Theta(1)$, if $\lambda<\tilde{\lambda}$, then for any $t\in[\tau_2,\tau_3]$,*

$$\forall i\in\mathcal{I}, \forall k\in[n], \langle \mathsf{w}_i^t, x_k\rangle > \delta'\|x_k\|.$$

Proof. Let

$$\mathcal{K}_{\delta'} \coloneqq \left\{k\in[n] \mid \exists v\in\mathbb{R}^d, \langle v, \frac{x_k}{\|x_k\|}\rangle = \delta' \text{ and } \forall k'\in[n], \langle v, \frac{x_{k'}}{\|x_{k'}\|}\rangle \geq \delta'\right\}.$$

By continuity, the condition $\langle \mathsf{w}_i^t, \frac{x_k}{\|x_k\|}\rangle > \delta'$ would necessarily break for some $k\in\mathcal{K}_{\delta'}$ first. It is thus sufficient to show Lemma 18 for any $k\in\mathcal{K}_{\delta'}$. Since for all $i,j\in\mathcal{I}$ and $t\in[\tau_2,\tau_3]$, $\langle \mathsf{w}_i^t, \mathsf{w}_j^t\rangle\geq 1-\lambda^\varepsilon$ by definition of τ_3, we can use a decomposition similar to Lemma 16:

$$\begin{gathered}\mathsf{w}_j^t = \alpha_{ij}^t \mathsf{w}_i^t + v_{ij}^t \quad \text{where } \alpha_{ij}^t = 1+\mathcal{O}(\lambda^\varepsilon),\\ v_{ij}^t \perp \mathsf{w}_i^t \quad \text{and} \quad \|v_{ij}^t\| = \mathcal{O}(\lambda^\varepsilon),\end{gathered}$$

where we used the fact that $\|\mathsf{w}_i^t\|\geq 1-\lambda^\varepsilon$. Using this decomposition, we can write

$$\begin{aligned}h_{\theta^t}(x_k) &= \sum_{j\in\mathcal{I}} \alpha_{ij}^t (a_j^t)^2\langle \mathsf{w}_i^t, x_k\rangle + \sum_{j\in\mathcal{I}}(a_j^t)^2\langle v_{ij}^t, x_k\rangle + \sum_{i\in\mathcal{N}} a_i^t\langle w_i^t, x_k\rangle_+\\ &\leq A\langle \mathsf{w}_i^t, x_k\rangle + \mathcal{O}(\lambda^\varepsilon),\end{aligned}$$

where $A = \Theta(1)$ is a constant bounding $\sum_{j\in\mathcal{I}}(a_j^t)^2$, thanks to Lemma 17. From there,

$$\begin{aligned}\langle D^t, x_k\rangle &= \frac{1}{n}\sum_{k'\neq k}\left(y_{k'} - h_{\theta^t}(x_{k'})\right)\langle x_{k'}, x_k\rangle + \frac{1}{n}(y_k - h_{\theta^t}(x_k))\|x_k\|^2 \\ &\geq -\sqrt{\frac{1}{n}\sum_{k'=1}^{n}\left(y_{k'} - h_{\theta^t}(x_{k'})\right)^2}\sqrt{\frac{1}{n}\sum_{k'\neq k}\langle x_{k'}, x_k\rangle^2} + \frac{y_k}{n}\|x_k\|^2 - \frac{A}{n}\langle \mathsf{w}_i^t, x\rangle - \mathcal{O}\left(\lambda^{\varepsilon}\right).\end{aligned}$$

Using the non-increasing property of gradient flow,

$$\begin{aligned}\frac{1}{n}\sum_{k'=1}^{n}\left(y_{k'} - h_{\theta^t}(x_{k'})\right)^2 &\leq \frac{1}{n}\sum_{k'=1}^{n}\left(y_{k'} - h_{\theta^0}(x_{k'})\right)^2 \\ &= \frac{1}{n}\sum_{k'=1}^{n} y_k^2 + \mathcal{O}\left(\lambda^2\right).\end{aligned}$$

Note that $\mathcal{K}_{\delta'} \subset \mathcal{K}$, using the correspondence $w = v - \delta'\frac{x_k}{\|x_k\|}$ in the definition of both sets.[18] Thanks to Assumption 5, the constant β defined in Equation (46) is positive. This then implies for any $k \in \mathcal{K}_{\delta'}$ and $t \in [\tau_2, \tau_3]$:

$$\langle D^t, x_k\rangle \geq \beta - \frac{A}{n}\langle \mathsf{w}_i^t, x\rangle - \mathcal{O}\left(\lambda^{\varepsilon}\right).$$

Again by our choice of λ, it also comes by monotonicity of the loss

$$\|D^t\| \leq \sqrt{\frac{1}{n}\sum_{k'=1}^{n} y_{k'}^2}\sqrt{\frac{1}{n}\sum_{k'=1}^{n}\|x_{k'}\|^2} + \mathcal{O}\left(\lambda\right) = \overline{D} = \Theta(1).$$

From these last two inequalities, for any $k \in \mathcal{K}_{\delta'}$ and $t \in [\tau_2, \tau_3]$, as long as $\langle \mathsf{w}_i^t, x_{k'}\rangle \geq \delta'\|x_{k'}\|$ for any k'

$$\begin{aligned}\frac{\mathrm{d}\langle \mathsf{w}_i^t, x_k\rangle}{\mathrm{d}t} &= \langle D^t, x_k\rangle - \langle \mathsf{w}_i^t, D^t\rangle\langle \mathsf{w}_i^t, x_k\rangle \\ &\geq \beta - (\frac{A}{n} + \overline{D})\langle \mathsf{w}_i^t, x_k\rangle - \mathcal{O}\left(\lambda^{\varepsilon}\right) \\ &\geq \beta - \mathcal{O}\left(\langle \mathsf{w}_i^t, x_k\rangle\right) - \mathcal{O}\left(\lambda^{\varepsilon}\right).\end{aligned}$$

Now note that for small enough $\tilde{\lambda} = \Theta(1)$, we can choose $\delta' = \Omega(1)$ small enough, so that $\beta - (\frac{A}{n} + \overline{D})\delta'\|x_k\| - \mathcal{O}\left(\lambda^{\varepsilon}\right) > 0$ and that $\langle \mathsf{w}_i^{\tau_2}, x_k\rangle > \delta'\|x_k\|$, thanks to Lemma 5. This necessarily implies that $\langle \mathsf{w}_i^t, x_k\rangle > \delta'\|x_k\|$ for any $t \in [\tau_2, \tau_3]$, as $\langle \mathsf{w}_i^t, x_k\rangle$ would be increasing before crossing the $\langle \mathsf{w}_i^t, x_k\rangle = \delta'\|x_k\|$ threshold. □

Lemma 19. *There exist $\varepsilon_2^* = \Theta(1)$ and $\tilde{\varepsilon}_2^* = \Theta(1)$ small enough such that for ε_2 satisfying the conditions of Lemmas 5 and 6 and any $t \in [\tau_2, \tau_3]$,*

$$\sum_{i\in\mathcal{I}}(a_i^t)^2 \geq \varepsilon_2.$$

18. The case $w = \mathbf{0}$ is a particular case, which implies $v = \delta'\frac{x_k}{\|x_k\|}$ and so all the x_k would be aligned, which contradicts the fact that X is full rank.

Proof. Note that we can chose ε_2^* small enough, so that

$$\sum_{i\in\mathcal{I}}(a_i^t)^2 \le 2\varepsilon_2 \implies \forall k\in[n], y_k > h_{\theta^t}(x_k). \tag{52}$$

Moreover we have for any $t\in[\tau_2,\tau_3]$ the ODE

$$\begin{aligned}\frac{\mathrm{d}\sum_{i\in\mathcal{I}}(a_i^t)^2}{\mathrm{d}t} &= 2\sum_{i\in\mathcal{I}}(a_i^t)^2\langle \mathsf{w}_i^t, D^t\rangle \\ &= \frac{2}{n}\sum_{i\in\mathcal{I}}(a_i^t)^2\left(\sum_{k=1}^n\langle \mathsf{w}_i^t, x_k\rangle\left(y_k - h_{\theta^t}(x_k)\right)\right).\end{aligned}$$

Thanks to Lemma 18, all the terms $\langle \mathsf{w}_i^t, x_k\rangle$ in the sum are positive for $t\in[\tau_2,\tau_3]$. Thanks to Equation (52), this sum is thus positive as soon as $\sum_{i\in I}(a_i^t)^2 \le 2\varepsilon_2$. Thus during this phase, $\sum_{i\in\mathcal{I}}(a_i^t)^2$ is increasing as soon as it is smaller than $2\varepsilon_2$. By continuity and since it is exactly ε_2 at the beginning of the phase, it is always at least ε_2 during this phase. □

F.4 Phase 3

Define for this proof

$$\beta^t \coloneqq \sum_{i\in\mathcal{I}_t} a_i^t w_i^t \text{ and } R_t \coloneqq \frac{1}{2}\sum_{i\in\mathcal{N}_t}\|w_i^t\|^2$$
$$\text{where } \mathcal{I}_t \coloneqq \left\{ i\in[m] \middle| \exists k\in[n], \langle w_i^t, x_k\rangle > 0\right\}$$
$$\text{and } \mathcal{N}_t \coloneqq \left\{ i\in[m] \middle| \exists k,k'\in[n] \text{ s.t. } \langle w_i^t, x_k\rangle > 0 \text{ and } \langle w_i^t, x_{k'}\rangle < 0\right\}.$$

Note the β^t differs from $\beta_{\mathcal{I}}^t$ in the proof of Lemma 6, as it does not only count the neurons in $\mathcal{I}$, but the neurons in $\mathcal{I}_t \supset I$.

Lemma 20. *For any $t\in[\tau_3,\tau_4)$ and $x\in\mathbb{R}^2$,*

$$|h_{\theta^t}(x) - \langle\beta^*, x\rangle_+| = \mathcal{O}\left((\varepsilon_3+\varepsilon_4)\|x\|\right).$$

Proof. Thanks to Lemma 6,

$$|h_{\theta^{\tau_3}}(x) - \langle\beta^*, x\rangle_+| = \mathcal{O}\left((\varepsilon_3+\lambda^\varepsilon)\|x\|\right).$$

Moreover for $x \neq \mathbf{0}$,

$$\begin{aligned}
\frac{1}{\|x\|}|h_{\theta^t}(x) - h_{\theta^{\tau_3}}(x)| &\leq \sum_{i=1}^m \|a_i^t w_i^t - a_i^{\tau_3} w_i^{\tau_3}\|_2 \\
&\leq \sum_{i=1}^m |a_i^{\tau_3}| \|w_i^t - w_i^{\tau_3}\|_2 + |a_i^t - a_i^{\tau_3}| \|w_i^t\|_2 \\
&\leq \sum_{i=1}^m |a_i^{\tau_3}| \|w_i^t - w_i^{\tau_3}\|_2 + |a_i^t - a_i^{\tau_3}| |a_i^t| \\
&\leq 2 \max\left(\sqrt{\sum_{i=1}^m (a_i^{\tau_3})^2}, \sqrt{\sum_{i=1}^m (a_i^t)^2} \right) \|\theta^t - \theta^{\tau_3}\|_2 \\
&\leq 2 \left(\sqrt{\sum_{i=1}^m (a_i^{\tau_3})^2} + \varepsilon_4 \right) \varepsilon_4.
\end{aligned}$$

Lemma 17 yields

$$\sum_{i=1}^m (a_i^{\tau_3})^2 = \mathcal{O}(1),$$

so that

$$|h_{\theta^t}(x_k) - h_{\theta^{\tau_3}}(x_k)| = \mathcal{O}(\varepsilon_3 + \varepsilon_4 + \lambda^\varepsilon).$$

This finally yields Lemma 20 as $\lambda^\varepsilon = \mathcal{O}(\varepsilon_4)$. □

Note that for any $k \in [n]$, $\langle \beta^*, x_k \rangle_+ = \langle \beta^*, x_K \rangle$ in our example.

Consider in the following some $\delta_4' > 0$ such that Assumption 6 holds with $\delta_u \geq \delta_4'$ for any considered u. We then define

$$\alpha_4 \coloneqq \inf_{\substack{u \in A(\mathbb{R}^d) \\ \exists k, u_k = 1 \\ \exists k', u_{k'} = 1}} \inf_{\substack{k \in \mathcal{K}, u_k \neq 0 \\ \exists w \in A^{-1}(u) \cap \mathbb{S}_d, |\langle w, \frac{x_k}{\|x_k\|} \rangle| \leq \delta_4'}} \langle \tilde{D}_u^{\beta^*}, x_k \rangle. \tag{53}$$

Assumption 6 directly yields $\alpha_4 > 0$ since the infimum is over a finite set. We then consider a small enough $\tilde{\delta}_4 = \Theta(\min(\delta_4', \alpha_4))$ and define

$$c_4 \coloneqq \inf_{\substack{u \in A(\mathbb{R}^d) \\ \exists k, u_k = 1 \\ \exists k', u_{k'} = 1}} \inf_{w \in A_{\tilde{\delta}_4}^{-1}(u)} \langle \tilde{D}_u^{\beta^*}, \frac{w}{\|w\|} \rangle. \tag{54}$$

Here again, Assumption 6 yields $c_4 > 0$.

Lemma 21. *Under Assumptions 4, 5 and 6, if $\lambda, \varepsilon_3, \varepsilon_4$ and δ_4 satisfy the conditions of Lemma 7 for small enough constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^*$, then*

$$R_t \leq \mathcal{O}\left(\lambda^{2(1-\varepsilon)} e^{-c_4(t-\tau_3)} \right) \text{ for any } t \in [\tau_3, \tau_4).$$

Importantly in Lemma 21, the $\mathcal{O}$ hides a constant that only depends on the data (and not on t).

Proof. First not that for any $i \in \mathcal{N}_t$, $a_i^t < 0$. Moreover, for $u_i^t = A(w_i^t)$ and $\tilde{D}_u^{\beta^*}$ defined by in Equation (11).

$$\begin{aligned}\frac{\mathrm{d}\|w_i^t\|^2}{\mathrm{d}t} &= 2a_i^t\langle D(w_i^t,\theta^t), w_i^t\rangle \\ &\le 2a_i^t\langle \tilde{D}^{\beta^*}_{u_i^t}, w_i^t\rangle - 2a_i^t\|w_i^t\|\|\tilde{D}^{\beta^*}_{u_i^t} - \tilde{D}_{u_i^t}(\theta^t)\|,\end{aligned}$$

where $\tilde{D}_{u_i^t}(\theta^t) = \frac{1}{n}\sum_{k=1}^n \mathbb{1}_{u_k=1}(y_k - h_{\theta^t}(x_k))x_k$. Thanks to Lemma 20 for our choice λ,

$$\|\tilde{D}^{\beta^*}_{u_i^t} - \tilde{D}_{u_i^t}(\theta^t)\| = \mathcal{O}\left(\varepsilon_3 + \varepsilon_4\right)$$

The previous inequality becomes

$$\frac{\mathrm{d}\|w_i^t\|^2}{\mathrm{d}t} \le 2a_i^t\langle \tilde{D}^{\beta^*}_{u_i^t}, w_i^t\rangle - 2a_i^t\|w_i^t\|\mathcal{O}\left(\varepsilon_3 + \varepsilon_4\right). \tag{55}$$

First, $i \in \mathcal{N}_t$ implies that there are k, k' such that $u_{i\ k}^t = -1$ and $u_{i\ k'}^t = 1$. Moreover at time t, we have two possibilities:

- either $w_i^t \in A^{-1}_{\tilde{\delta}_4}(u_i^t)$;
- or $w_i^t \notin A^{-1}_{\tilde{\delta}_4}(u_i^t)$.

In the former, Equation (55) yields for small enough $\varepsilon_3^*, \varepsilon_4^* = \Theta(1)$

$$\begin{aligned}\frac{\mathrm{d}\|w_i^t\|^2}{\mathrm{d}t} &\le 2c_4a_i^t\|w_i^t\| - 2a_i^t\|w_i^t\|\mathcal{O}\left(\varepsilon_3 + \varepsilon_4\right) \\ &\le -(2c_4 - \mathcal{O}\left(\varepsilon_3 + \varepsilon_4\right))\|w_i^t\|^2.\end{aligned} \tag{56}$$

In the latter case, by definition of $A^{-1}_{\tilde{\delta}_4}$, we either have $\langle \frac{w_i^t}{\|w_i^t\|}, \frac{x_k}{\|x_k\|}\rangle \le \tilde{\delta}_4$ for all $k \in \mathcal{K}$ or $-\langle \frac{w_i^t}{\|w_i^t\|}, \frac{x_k}{\|x_k\|}\rangle \le \tilde{\delta}_4$ for all $k \in \mathcal{K}$. Assume in the following the first case and consider $k \in \mathcal{K}$ such that $\langle \frac{w_i^t}{\|w_i^t\|}, \frac{x_k}{\|x_k\|}\rangle \le \tilde{\delta}_4$. The symmetric case is dealt with similarly.

Denote[19] in the following $\tilde{\mathsf{w}}_i^t = \frac{w_i^t}{\|w_i^t\|}$. From there, it comes

$$\frac{\mathrm{d}\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|}\rangle}{\mathrm{d}t} \in \frac{a_i^t}{\|w_i^t\|}\left(\langle \mathfrak{D}_i^t, \frac{x_k}{\|x_k\|}\rangle - \langle \mathfrak{D}_i^t, \tilde{\mathsf{w}}_i^t\rangle\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|}\rangle\right).$$

Thanks to Assumption 6 and the definition of $\tilde{\delta}_4$

$$\begin{aligned}\frac{\mathrm{d}\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|}\rangle}{\mathrm{d}t} &\le \frac{a_i^t}{\|w_i^t\|}\left(\alpha_4 - \overline{D}\tilde{\delta}_4 - \mathcal{O}\left(\varepsilon_3 + \varepsilon_4\right)\right) \\ &\le \frac{a_i^t}{\|w_i^t\|}\left(\alpha_4 - \mathcal{O}\left(\tilde{\delta}_4 + \varepsilon_3 + \varepsilon_4\right)\right)\end{aligned}$$

From there, we can choose $\varepsilon_3^*, \varepsilon_4^*$ and $\tilde{\delta}_4$ small enough (but still $\Theta(1)$) so that $\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|}\rangle$ decreases at a rate $\frac{a_i^t}{\|w_i^t\|}\Theta(\alpha_4)$. This reasoning actually holds for any $k \in \mathrm{argmax}_{k\in\mathcal{K}}\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|}\rangle$

19. This can be defined, as $w_i^t \neq \mathbf{0}$ as long as $i \notin N_t$.

as long as $i \in N_t$ and $w_i^t \notin A_{\delta_4}^{-1}(u_i^t)$. As a consequence, if all $\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|} \rangle$ are smaller than δ_4 with at least one of them positive for $k \in \mathcal{K}$, then their maximal value (among $\mathcal{K}$) decreases at the above rate, so that after a time at most $\Theta(1)$, all the values for $k \in \mathcal{K}$ become non-negative for $k \in \mathcal{K}$. This would then also imply that all values $\langle \tilde{\mathsf{w}}_i^t, \frac{x_k}{\|x_k\|} \rangle$ for $k \in [n]$ become non-negative. Indeed, if $\{k' \mid \langle w, x_{k'} \rangle > 0\}$ is non-empty, we can, thanks to Assumption 4, substract vectors of the form $\alpha_{k'} x_{k'}$ to w. This would decrease all values $\langle w, x_{k'} \rangle$. We can proceed this strategy until all the scalar products are non-positive and at least one of them is[20] 0. This would then imply that the k' for which it is 0 are in $\mathcal{K}$, i.e., $\{k' \mid \langle w, x_{k'} \rangle > 0\}$ is either empty, or intersects $\mathcal{K}$.

So if at some point $i \in \mathcal{N}_t$ and $w_i^t \notin A_{\tilde{\delta}_4}^{-1}(u_i^t)$, then after a time Δt , $i \notin \mathcal{N}_{t+\Delta t}$ where Δt satisfies

$$\int_t^{t+\Delta t} \frac{|a_i^s|}{\|w_i^s\|} \mathrm{d}s = \Theta\left(1\right). \tag{57}$$

Moreover, during this whole time,

$$\frac{\mathrm{d}\|w_i^t\|^2}{\mathrm{d}t} = \mathcal{O}\left(\frac{|a_i^t|}{\|w_i^t\|}\|w_i^t\|^2\right). \tag{58}$$

Also, the above argument yields that the sets $\mathcal{N}_t$ are non-increasing over time on $[\tau_3, \tau_4]$. Indeed, a neuron i cannot enter in the set $\mathcal{N}_t$, as it would either have $w_i^t \neq \mathbf{0}$ and $\langle \tilde{\mathsf{w}}_i^t, x_k \rangle = 0$ for some $k \in \mathcal{K}$ at entrance, or $w_i^t = \mathbf{0}$ at entrance. In the first case, $w_i^t \notin A_{\tilde{\delta}_4}^{-1}(u_i^t)$ and is thus immediately ejected out of $\mathcal{N}_t$ (and is hence not entering). In the second case, note that $\frac{a_i^t}{\|w_i^t\|}$ is arbitrarily large at entrance in $\mathcal{N}_t$. Thus, if w_i^t enters such that $w_i^t \notin A_{\tilde{\delta}_4}^{-1}(u_i^t)$, it is also immediately ejected out[21] of $\mathcal{N}_t$. If instead w_i^t enters such that $w_i^t \in A_{\tilde{\delta}_4}^{-1}(u_i^t)$, its norm decreases following Equation (56), while it is actually 0 at entrance. Hence in all the cases, it cannot enter $\mathcal{N}_t$.

For the neuron $i \in \mathcal{N}_{\tau_3}$, we can now partition $[\tau_3, \tau_4)$ into three successive disjoint intervals such that

- Equation (56) holds in $[\tau_3, t_0]$;
- the second interval is of the form $[t_0, t_0 + \Delta t]$ with Equations (57) and (58);
- $i \notin \mathcal{N}_t$ in $[t_0 + \Delta t, \tau_4)$.

Using this partition, a Grönwall argument then yields for any $t \in [\tau_3, \tau_4)$ and $\varepsilon_3^*, \varepsilon_4^*$ small enough,

$$\begin{aligned}\|w_i^t\|^2 \mathbb{1}_{i \in \mathcal{N}_t} &\leq \mathbb{1}_{t \leq t_0 + \Delta t} \|w_i^{\tau_3}\|^2 e^{-\frac{c_4}{2}(\min(t, t_0) - \tau_3)} \exp\left(\mathcal{O}\left(\int_{t_0}^{t_0+\Delta t} \frac{|a_i^s|}{\|w_i^s\|} \mathrm{d}s\right)\right) \\ &= \mathcal{O}\left(\|w_i^{\tau_3}\|^2 e^{-\frac{c_4}{2}(t-\tau_3)}\right).\end{aligned}$$

20. The vector resulting from these substractions cannot be $\mathbf{0}$, since it would imply that w is positively correlated with all data points.

21. A simpler argument is here to directly consider the non-scaled version $\frac{\mathrm{d}\langle w_i^t, x_k \rangle}{\mathrm{d}t}$.

From there, using the fact that $R_{\tau_3} \leq \lambda^{2(1-\varepsilon)}$, a simple summation yields

$$R_t = \mathcal{O}\left(\lambda^{2(1-\varepsilon)} e^{-\frac{c_4}{2}(t-\tau_3)}\right) \text{ for any } t \in [\tau_3, \tau_4).$$

□

Lemma 22. *Under Assumptions 4, 5 and 6, if $\lambda, \varepsilon_3, \varepsilon_4$ and δ_4 satisfy the conditions of Lemma 7 for small enough constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^*$, then for any $t \in (\tau_3, \tau_4)$:*

$$\forall g_t \in \partial L(\theta^t), \|g_t\|^2 \geq \sigma_{\min}(H)\|\beta^*\| \left(L(\theta^t) - L_{\beta^*}\right) + \mathcal{O}\left(\lambda^{2(1-\varepsilon)} e^{-\frac{c_4}{4}(t-\tau_3)}\right),$$

where $L_{\beta^*} = \frac{1}{2n}\sum_{k=1}^n \left(\langle \beta^*, x_k\rangle - y_k\right)^2$ and $\sigma_{\min}(H)$ is the smallest eigenvalue of $H = \frac{1}{n} X^\top X$.

Proof. For any $t \in (\tau_3, \tau_4)$, it comes by definition of τ_4 that $\langle w_i^t, x_k\rangle > 0$ for any $i \in \mathcal{I}$ and $k \in [n]$. It then holds for any $g_t \in \partial L(\theta^t)$, when considering only the norm of its components corresponding to the derivatives along w_i^t for $i \in \mathcal{I}$:

$$\|g_t\|_2^2 \geq \sum_{i \in \mathcal{I}} (a_i^t)^2 \|D^t\|^2. \tag{59}$$

Note that by definition of β^t and R_t:

$$\begin{aligned} h_{\theta^t}(x_k) &= \sum_{i=1}^m a_i^t \langle w_i^t, x_k\rangle_+ \\ &= \sum_{i=1}^m a_i^t (\langle w_i^t, x_k\rangle + \langle w_i^t, x_k\rangle_-) \\ &= \langle \beta^t, x_k\rangle + \sum_{i \in \mathcal{N}_t} a_i^t \langle w_i^t, x_k\rangle_- \ . \end{aligned}$$

So that

$$\begin{aligned} |h_{\theta^t}(x_k) - \langle \beta^t, x_k\rangle| &\leq \sum_{i \in \mathcal{N}_t} |a_i^t| \|w_i^t\| \|x_k\| \\ &\leq \sqrt{\sum_{i \in \mathcal{N}_t} (a_i^t)^2} \sqrt{\sum_{i \in \mathcal{N}_t} \|w_i^t\|^2} \|x_k\| \\ &\leq \sqrt{2R_t + \lambda^2}\sqrt{2R_t}\|x_k\| \\ &= \mathcal{O}\left(\lambda^{2(1-\varepsilon)} e^{-\frac{c_4}{4}(t-\tau_3)}\right). \end{aligned}$$

The last inequality here comes from Lemma 21. By definition of D^t and β^*,

$$\begin{aligned}
D^t &= -\frac{1}{n}\sum_{k=1}^{n}(h_{\theta^t}(x_k) - y_k)x_k \\
&= -\frac{1}{n}\sum_{k=1}^{n}(h_{\theta^t}(x_k) - \langle\beta^*, x_k\rangle)x_k \\
&= -H(\beta^t - \beta^*) - \frac{1}{n}\sum_{k=1}^{n}(h_{\theta^t}(x_k) - \langle\beta^t, x_k\rangle)x_k \\
&= -H(\beta^t - \beta^*) + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right).
\end{aligned}$$

In the following, we note $\sigma_{\min}$ instead of $\sigma_{\min}(H)$ for simplicity. Equation (59) now becomes for $t \in (\tau_3, \tau_4)$

$$\begin{aligned}
\|g_t\|_2^2 &\geq \sum_{i\in\mathcal{I}}(a_i^t)^2\left(\sqrt{\sigma_{\min}}\|\beta^t - \beta^*\|_H - \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right)\right)^2 \\
&\geq \sigma_{\min}\sum_{i\in\mathcal{I}}(a_i^t)^2\|\beta^t - \beta^*\|_H^2 - \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right) \\
&\geq (\|\beta^*\|_2 - \mathcal{O}(\varepsilon_3 + \varepsilon_4))\sigma_{\min}\|\beta^t - \beta^*\|_H^2 - \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right), \qquad (60)
\end{aligned}$$

where the second inequality uses a bound $\|\beta^t - \beta^*\|_2 = \mathcal{O}(1)$, which can be proved similarly to Lemma 20. Note that we again have $\sum_{i\in\mathcal{I}}(a_i^t)^2 = \mathcal{O}(1)$ in this phase, thanks to Lemma 17 and the definition of τ_4.

On the other hand, we also have for $\tilde{R}_t(x) \coloneqq h_{\theta^t}(x) - \langle\beta^t, x\rangle$

$$\begin{aligned}
L(\theta^t) - L_{\beta^*} &= \frac{1}{2n}\sum_{k=1}^{n}(h_{\theta^t}(x_k) - y_k)^2 - (\langle\beta^*, x_k\rangle - y_k)^2 \\
&= \frac{1}{2n}\sum_{k=1}^{n}(\langle\beta^t, x_k\rangle - y_k)^2 - (\langle\beta^*, x_k\rangle - y_k)^2 + \frac{1}{2n}\sum_{k=1}^{n}(2(\langle\beta^t, x_k\rangle - y_k)\tilde{R}_t(x_k) + \tilde{R}_t(x_k)^2) \\
&= \frac{1}{n}(\beta^t - \beta^*)^\top X^\top(X\beta^* - Y) + \frac{1}{2n}(\beta^t - \beta^*)^\top X^\top X(\beta^t - \beta^*) \\
&\quad + \frac{1}{2n}\sum_{k=1}^{n}\left(2(h_{\theta^t}(x_k) - y_k)\tilde{R}_t(x_k) - \tilde{R}_t(x_k)^2\right) \\
&\leq \frac{1}{2}\|\beta^t - \beta^*\|_H^2 + \frac{1}{n}\sum_{k=1}^{n}(h_{\theta^t}(x_k) - y_k)\tilde{R}_t(x_k).
\end{aligned}$$

The last inequality comes from the definition of β^* that implies $X^\top(X\beta^* - Y) = \mathbf{0}$. Recall that $|\tilde{R}_t(x_k)| = \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right)$. Also, note that $\frac{1}{n}\sum_{k=1}^{n}(h_{\theta^t}(x_k) - y_k) = \mathcal{O}(1)$ by monotonicity of the loss. It yields with Lemma 21:

$$L(\theta^t) - L_{\beta^*} \leq \frac{1}{2}\|\beta^t - \beta^*\|_H^2 + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right). \qquad (61)$$

Combining Equations (60) and (61); it finally yields:

$$\begin{aligned}\|g_t\|_2^2 &\geq \sigma_{\min}\left(\|\beta^*\| - \mathcal{O}\left(\varepsilon_3+\varepsilon_4\right)\right)\|\beta^t-\beta^*\|_H^2 - \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right)\\ &\geq 2\sigma_{\min}\left(\|\beta^*\| - \mathcal{O}\left(\varepsilon_3+\varepsilon_4\right)\right)\left(L(\theta^t)-L_{\beta^*}\right) - \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right).\end{aligned}$$

Lemma 22 then follows for small enough ε_3^* and ε_4^*, so that the first term is larger than $\sigma_{\min}\|\beta^*\|\left(L(\theta^t)-L_{\beta^*}\right)$. □

Lemma 23. *Under Assumptions 4, 5 and 6, if $\lambda, \varepsilon_3, \varepsilon_4$ and δ_4 satisfy the conditions of Lemma 7 for small enough constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^*$, then for any $t \in [\tau_3, \tau_4)$ and $a \coloneqq \min\left(\frac{\sigma_{\min}(H)\|\beta^*\|}{2}, \frac{c_4}{4}\right)$:*

$$L(\theta^t) - L_{\beta^*} = \mathcal{O}\left(\varepsilon_3^2 e^{-a(t-\tau_3)}\right).$$

Proof. By definition of the gradient flow, there is some $g_t \in \partial L(\theta^t)$ such that $\frac{\mathrm{d}\theta^t}{\mathrm{d}t} = -g_t$ a.e. It then comes from Lemma 22 that a.e.[22]

$$\begin{aligned}\frac{\mathrm{d}(L(\theta^t)-L_{\beta^*})}{\mathrm{d}t} &= -\|g_t\|^2\\ &\leq -\sigma_{\min}\|\beta^*\|(L(\theta^t)-L_{\beta^*}) + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-\frac{c_4}{4}(t-\tau_3)}\right)\\ &\leq -\sigma_{\min}\|\beta^*\|(L(\theta^t)-L_{\beta^*}) + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-a(t-\tau_3)}\right),\end{aligned}$$

where again we write $\sigma_{\min}$ for $\sigma_{\min}(H)$. Solutions of the ODE $f'(t) = -cf(t) + be^{-at}$ are of the form

$$f(t) = \frac{b}{c-a}(e^{-at}-e^{-ct}) + f(0)e^{-ct} \quad \text{if } a \neq c.$$

Since $a < \sigma_{\min}\|\beta^*\|$, a Grönwall comparison then yields for any $t \in [\tau_3, \tau_4)$,

$$L(\theta^t) - L_{\beta^*} \leq \left(L(\theta^{\tau_3}) - L_{\beta^*}\right)e^{-\sigma_{\min}\|\beta^*\|(t-\tau_3)} + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}e^{-a(t-\tau_3)}\right).$$

Moreover, it comes from Equation (61) that

$$\begin{aligned}L(\theta^{\tau_3}) - L_{\beta^*} &\leq \frac{1}{2}\|\beta^{\tau_3}-\beta^*\|_H^2 + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}\right)\\ &\leq \frac{\sigma_{\max}}{2}\|\beta^{\tau_3}-\beta^*\|_2^2 + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}\right)\\ &\leq \frac{\sigma_{\max}}{2}\varepsilon_3^2 + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}\right),\end{aligned}$$

where the last inequality comes from Lemma 6. This allows to conclude as $\lambda^{1-\varepsilon} = \mathcal{O}(\varepsilon_3)$. □

Lemma 24. *Under Assumptions 4, 5 and 6, if $\lambda, \varepsilon_3, \varepsilon_4$ and δ_4 satisfy the conditions of Lemma 7 for small enough constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^*$, then*

$$\int_{\tau_3}^{\tau_4}\|D^t\|\mathrm{d}t = \mathcal{O}\left(\varepsilon_3\right).$$

22. The fact that the chain rule can be applied to $\frac{\mathrm{d}L(\theta^t)}{\mathrm{d}t}$ here is not straightforward, but is possible a.e. (see e.g., Bolte and Pauwels, 2021, Lemma 2, Corollary 1 and Proposition 2)

Proof. We already proved in the proof of Lemma 22 for any $t \in [\tau_3, \tau_4)$:

$$\|D_t\| \leq \sqrt{\sigma_{\max}}\|\beta^t - \beta^*\|_H + \mathcal{O}\left(\lambda^{2(1-\varepsilon)} e^{-\frac{c_4}{4}(t-\tau_3)}\right).$$

Moreover, we also showed

$$\|\beta^t - \beta^*\|_H^2 = 2(L(\theta^t) - L_{\beta^*}) + \frac{1}{n}\sum_{k=1}^{n}\left((y_k - h_{\theta^t}(x_k))\tilde{R}_t(x_k) + \tilde{R}_t(x_k)^2\right)$$
$$\|\beta^t - \beta^*\|_H \leq \sqrt{2(L(\theta^t) - L_{\beta^*})} + \mathcal{O}\left(\lambda^{(1-\varepsilon)} e^{-\frac{c_4}{4}(t-\tau_3)}\right).$$

Thanks to Lemma 23, it comes for our choice of λ and ε_3:

$$\|D_t\| \leq \mathcal{O}\left(\varepsilon_3 e^{-\frac{a}{2}(t-\tau_3)}\right),$$

where the constants hidden in $\mathcal{O}$ do not depend on t but only the dataset. Integrating the exponential then concludes the proof. □

Lemma 25. *Under Assumptions 4, 5 and 6, if $\lambda, \varepsilon_3, \varepsilon_4$ and δ_4 satisfy the conditions of Lemma 7 for small enough constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^*$, then for any $i \in \mathcal{I}$ and $t \in [\tau_3, \tau_4)$:*

- $|a_i^t - a_i^{\tau_3}| \leq \left(e^{\mathcal{O}(\varepsilon_3)} - 1\right) a_i^{\tau_3}$;
- $\|w_i^t - w_i^{\tau_3}\| \leq \left(e^{\mathcal{O}(\varepsilon_3)} + \mathcal{O}(\varepsilon_3) - 1\right) a_i^{\tau_3}$;
- $\|\mathsf{w}_i^t - \mathsf{w}_i^{\tau_3}\| \leq \mathcal{O}(\varepsilon_3)$.

Moreover, for any $i \in \mathcal{N}$ and $t \in [\tau_3, \tau_4)$,

- $|a_i^t - a_i^{\tau_3}| \leq \mathcal{O}(1)\, a_i^{\tau_3}$;
- $\|w_i^t - w_i^{\tau_3}\| \leq \mathcal{O}(1)\, a_i^{\tau_3}$,

where again the constants hidden in $\mathcal{O}$ neither depend on t nor i, but only on the dataset.

Proof. For any $i \in \mathcal{I}$, the neuron i is activated along all x_k for all $t \in [\tau_3, \tau_4)$ by definition of τ_4. As a consequence, we can bound the derivative of $a_i^t > 0$ as:

$$\left|\frac{\mathrm{d}a_i^t}{\mathrm{d}t}\right| = |\langle w_i^t, D^t\rangle|$$
$$\leq a_i^t \|D^t\|.$$

Grönwall inequality along with Lemma 24 then yield the first item of Lemma 25. We also have the following bound:

$$\left\|\frac{\mathrm{d}\mathsf{w}_i^t}{\mathrm{d}t}\right\| = \left\|D^t - \langle \mathsf{w}_i^t, D^t\rangle \mathsf{w}_i^t\right\|$$
$$\leq \|D^t\|.$$

This yields the third point of Lemma 25. Moreover note that

$$\|w_i^t - w_i^{\tau_3}\| \leq |a_i^t - a_i^{\tau_3}| + a_i^{\tau_3}\|\mathsf{w}_i^t - \mathsf{w}_i^{\tau_3}\|.$$

The second point then follows from the first and third one.

For $i \in \mathcal{N}$, the neuron i might not be activated along all x_k. However, we can use the same arguments as in the proof of Lemma 21, so that there are time t_i (potentially ∞) and Δt_i, such that $|a_i^t|$ decreases on $[\tau_3, t_i]$ and i is activated along all (or no) x_k on $[t_i + \Delta t_i, \tau_4)$ with $\Delta t_i = \mathcal{O}(1)$. From the first part, it holds for any $t \in [\tau_3, t_i]$

$$|a_i^t - a_i^{\tau_3}| \leq |a_i^{\tau_3}|$$
$$\|w_i^t - w_i^{\tau_3}\| \leq \|w_i^t\| + \|w_i^{\tau_3}\| \leq 2|a_i^{\tau_3}|.$$

And a Grönwall argument also yields for any $t \in [\tau_3, t_i + \Delta t_i]$, since $\Delta t_i = \mathcal{O}(1)$:

$$|a_i^t - a_i^{\tau_3}| = \mathcal{O}\left(|a_i^{\tau_3}|\right)$$
$$\|w_i^t - w_i^{\tau_3}\| = \mathcal{O}\left(|a_i^{\tau_3}|\right).$$

Also, we can bound the difference in the third part similarly to the case $i \in \mathcal{I}$, i.e., for any $t \in [t_i + \Delta t_i, \tau_4)$:

$$|a_i^t - a_i^{t_i + \Delta t_i}| \leq a_i^{t_i + \Delta t_i} e^{\mathcal{O}(\varepsilon_3)}$$
$$\|w_i^t - w_i^{t_i + \Delta t_i}\| \leq \left(e^{\mathcal{O}(\varepsilon_3)} + \mathcal{O}(\varepsilon_3)\right) a_i^{t_i + \Delta t_i}.$$

Combining these different inequalities then yields the last two points of Lemma 25, using the fact that $\varepsilon_3 = \mathcal{O}(1)$. □

Corollary 1. *Under Assumption 6, if $\lambda, \varepsilon_3, \varepsilon_4$ and δ_4 satisfy the conditions of Lemma 7 for small enough constants $\tilde{\lambda}, \varepsilon_3^*, \varepsilon_4^*, \delta_4^*$, then $\tau_4 = \infty$.*

Proof. First note that thanks to Lemma 6 and Lemma 25, we have for any $t \in [\tau_3, \tau_4)$, $i \in \mathcal{I}$ and $k \in [n]$:

$$\begin{aligned}\langle \mathsf{w}_i^t, x_k \rangle &\geq \langle \mathsf{w}_i^{\tau_3}, x_k \rangle - \|\mathsf{w}_i^t - \mathsf{w}_i^{\tau_3}\| \, \|x_k\| \\ &\geq \Omega(1) - \mathcal{O}(\varepsilon_3) \\ &> \delta_4 \|x_k\|,\end{aligned}$$

for a small enough choice of ε_3^* and δ_4^*. This implies that the second condition in the definition of τ_4 does not break first. Moreover, thanks to Lemma 25,

$$\begin{aligned}\|\theta^t - \theta^{\tau_3}\|_2^2 &\leq \sum_{i \in I} (a_i^t - a_i^{\tau_3})^2 + \|w_i^t - w_i^{\tau_3}\|^2 + \sum_{i \in \mathcal{N}} (a_i^t - a_i^{\tau_3})^2 + \|w_i^t - w_i^{\tau_3}\|^2 \\ &= \left(e^{\mathcal{O}(\varepsilon_3)} + \mathcal{O}(\varepsilon_3) - 1\right)^2 \sum_{i \in \mathcal{I}} (a_i^{\tau_3})^2 + \mathcal{O}(1) \sum_{i \in \mathcal{N}} (a_i^{\tau_3})^2 \\ &\leq \left(e^{\mathcal{O}(\varepsilon_3)} + \mathcal{O}(\varepsilon_3) - 1\right)^2 \mathcal{O}(1) + \mathcal{O}\left(\lambda^{2(1-\varepsilon)}\right),\end{aligned}$$

where the last inequality comes from the bounds on the sum of $(a_i^{\tau_3})^2$ from Lemmas 6 and 19. Now for any $\varepsilon_4^* = \Theta(1)$ and small enough $\varepsilon_3^*, \tilde{\lambda} = \Theta(1)$ (depending on ε_4^*), the previous inequality leads to

$$\|\theta^t - \theta^{\tau_3}\|_2^2 < \varepsilon_4^*.$$

This implies that the first condition in the definition of τ_4 does not break first. As a consequence, neither of the conditions in the definition of τ_4 break in finite time, i.e., $\tau_4 = \infty$. □

Proof of Lemma 7. First note that we can choose constants $\varepsilon, \tilde{\lambda}, \varepsilon_2, \varepsilon_3, \varepsilon_4, \delta_4 = \Theta(1)$, such that all the lemmas of Appendix F simultaneously hold for any $\lambda \leq \tilde{\lambda}$. An easy way to verify so is going backward after fixing ε, i.e., first fix ε_4, δ_4, then fixing ε_3, then ε_2 and finally $\tilde{\lambda}$. Thanks to Lemmas 24 and 25 and Corollary 1, the parameters θ^t and their variations are both bounded on $[\tau_3, \infty)$. As a consequence, θ^t does have a limit:

$$\lim_{t\to\infty} \theta^t = \theta_\lambda^\infty.$$

Moreover, Lemma 21 yields that

$$\lim_{t\to\infty} R_t = 0.$$

This directly implies the second point of Lemma 7. Now denote

$$\beta_\lambda^\infty \coloneqq \sum_{\substack{i\in[m] \\ \forall k\in[m], \langle w_{\lambda,i}^\infty, x_k\rangle \geq 0}} a_{\lambda,i}^\infty w_{\lambda,i}^\infty.$$

The second point of Lemma 7 that we just proved implies that

$$h_{\theta^\infty}(x_k) = \langle \beta_\lambda^\infty, x\rangle \quad \text{for any } x \in \mathcal{C},$$

i.e. θ_λ^∞ behaves as a linear regression of parameter β_λ^∞ on the (cone generated by the) convex hull of the training data. As a consequence,

$$L(\theta_\lambda^\infty) = L_{\beta_\lambda^\infty} \coloneqq \frac{1}{2n}\sum_{k=1}^n \left(\langle \beta_\lambda^\infty, x_k\rangle - y_k\right)^2.$$

Also, Lemma 23 implies that $L(\theta_\lambda^\infty) = L_{\beta_\lambda^\infty} \leq L_{\beta^*}$. However, β^* is the unique minimiser of the least square loss, among the set of linear regression parameters. This implies that $\beta_\lambda^\infty = \beta^*$, which concludes the proof of both Lemma 7 and Theorem 2.

The last point of Theorem 4 is obtained by using Lemma 20 and noticing that when $\lambda \to 0$, we can also choose the constants ε_3 and ε_4 such that they converge to 0 when $\lambda \to 0$.

Appendix G. Further Dicussion on (Glasgow, 2024)

In this section, we discuss in more details how the early alignment phenomenon is related to Glasgow (2024)'s work. Glasgow (2024) studies the convergence of SGD for XOR-type data with two-layer ReLU network towards four vectors, resulting in a 0 test loss.

We argue that the XOR setting under consideration falls within the scope of our framework. Specifically, the associated population loss exhibits exactly four extremal vectors corresponding to these four vectors reached at convergence (see details below). From this, the initial phase of the dynamics described by Glasgow (2024) aligns with the early alignment phase analyzed in our work—Theorem 1 has yet to be applied to gradient flow over the population loss, which requires extending it to the infinite data setting.

The primary additional technical challenges addressed in Glasgow (2024) involve establishing analogous results for stochastic gradient descent (SGD) rather than gradient flow, and under a finite data regime. Their analysis of early alignment, and the broader training dynamics, for SGD in the XOR setting is achieved by bounding the deviation between SGD and gradient flow on the population loss via concentration inequalities. Controlling the growth of the neurons from the early alignment phase also remains challenging, although simpler to control.

G.1 Computation of Extremal Vectors in Glasgow 2024 Setting.

In this section, we justify that the population loss only counts four extremal vectors in the XOR setting. For that, we consider the *population loss* with Gaussian input.

While Glasgow (2024) consider data distributed uniformly over the hypercube, their proof of the first alignment phase relies on an approximation of the uniform distribution on the hypercube by a standard Gaussian in high dimension. In that effect, we directly assume here that the dataset is given by

$$\begin{aligned} x_k &\sim \mathcal{N}(0, \mathrm{I}_d) \\ y(x_k) &= -\mathrm{sign}(e_1^\top x_k)\mathrm{sign}(e_2^\top x_k) \end{aligned}$$

and consider the limit of infinite data. We thus denote the (population) loss of the model as

$$L(\theta) \coloneqq \mathbb{E}_{x,y}\left[\ell(h_\theta(x), y(x))\right],$$

where $\ell(\hat{y}, y) = \ln(1 + e^{-\hat{y}y})$ is the logistic loss. In this population loss setting, the function $G(w)$ can still be defined, and its gradient is given by

$$D(w, \mathbf{0}) = \mathbb{E}_{x,y}[yx\mathbb{1}_{w^\top x \geq 0}].$$

In that infinite data setting, the definition of extremal vector also extends to $D \neq \mathbf{0}$ such that both hold (see Boursier and Flammarion, 2024)

$$\begin{aligned} &1. \quad D = D(w, \mathbf{0}) \\ &2. \quad \frac{D}{\|D\|} = \pm\frac{w}{\|w\|}. \end{aligned}$$

Using computations (see Appendices G.2 and G.3 below) similar to Lemma D.4 (Glasgow, 2024), we can then show that for any $w \in \mathbb{R}^d \setminus \{0\}$:

$$\begin{aligned} \mathrm{sign}(w_2) &= -\mathrm{sign}(\langle D(w, \mathbf{0}), e_1\rangle), \\ \mathrm{sign}(w_1) &= -\mathrm{sign}(\langle D(w, \mathbf{0}), e_2\rangle) \end{aligned} \tag{62}$$

and[23]

$$\mathrm{sign}(\langle D(w, \mathbf{0}), w_{3:d}\rangle) = \mathrm{sign}(w_1 w_2)\mathbb{1}_{w_{3:d} \neq 0}. \tag{63}$$

From there, assume that $D(w, \mathbf{0})$ is an extremal vector. If $w_1 = 0$, then the above imply that $D(w, \mathbf{0}) \perp e_j$ for any $j \geq 2$, so that it is not an extremal vector. In consequence, we necessarily have $w_1 \neq 0$ and $w_2 \neq 0$ for similar reasons.

23. In this section, w_i is the i-th coordinate of w and $w_{i:j}$ is the projection of w onto $\mathrm{Span}(e_i, \dots, e_j)$.

From there assume w.l.o.g. that $w_1 w_2 > 0$ and also assume that $w_{3:d} \neq 0$. We then have by Equations (62) and (63):

$$\langle D(w, \mathbf{0}), w_{1:2} \rangle < 0,$$
$$\langle D(w, \mathbf{0}), w_{3:d} \rangle > 0.$$

However, these two inequalities contradict the fact that $D(w, \mathbf{0}) \propto \pm w$, i.e., that it is an extremal vector. Thus, any extremal vector $D(w, \mathbf{0})$ satisfies $w_{3:d} = 0$. Necessarily, it must be of the form $D(w, \mathbf{0}) = \alpha_1 e_1 + \alpha_2 e_2$. We can now be more precise and show that (see Appendix G.4 below) for any w such that $w_{3:d} = 0$,

$$\begin{aligned} &\text{if} \quad |w_1| > |w_2| \quad \text{then} \quad |\langle D(w, \mathbf{0}), e_2 \rangle| > |\langle D(w, \mathbf{0}), e_1 \rangle|, \\ &\text{if} \quad |w_1| < |w_2| \quad \text{then} \quad |\langle D(w, \mathbf{0}), e_2 \rangle| < |\langle D(w, \mathbf{0}), e_1 \rangle|. \end{aligned} \tag{64}$$

In consequence, any extremal vector must be such that $|w_1| = |w_2|$, i.e., of the form $D(w, \mathbf{0}) = \alpha(e_1 \pm e_2)$ for some $\alpha \in \mathbb{R}$. In consequence, there exist at most 4 extremal vectors and it is easy to check that for a good choice of α (both > 0 and < 0), this indeed yields extremal vectors.

We have thus shown that in the XOR setting with population logistic loss and Gaussian data, there are only 4 extremal vectors, which are proportional to the vectors $e_1 + e_2$, $e_1 - e_2$, $-e_1 + e_2$, $e_1 + e_2$.

G.2 Proof of Equation (62).

Similarly to Glasgow (2024), we note $x = z + \xi$ where z is the projection of x on the first two coordinates and ξ on the last $d - 2$ coordinates. We also note $y(z) = -\text{sign}(z_1 z_2)$. Denoting by $\text{swap}_1(x) = (-x_1, x_2, \ldots, x_d)$ the flip operator on the first coordinate, we have using the symmetries of the distribution:

$$\begin{aligned} \langle D(w, \mathbf{0}), e_2 \rangle &= \mathbb{E}[y(z) \mathbb{1}_{w^\top x \geq 0} x_2] \\ &= \frac{1}{2} \mathbb{E}[(\mathbb{1}_{w^\top x \geq 0} - \mathbb{1}_{w^\top \text{swap}_1(x) \geq 0}) y(z) x_2] \\ &= -\frac{1}{2} \mathbb{E}[(\mathbb{1}_{w^\top x \geq 0} - \mathbb{1}_{w^\top \text{swap}_1(x) \geq 0}) \text{sign}(x_1) |x_2|]. \end{aligned} \tag{65}$$

Moreover, note that if $w_1 \geq 0$, we necessarily have $(\mathbb{1}_{w^\top x \geq 0} - \mathbb{1}_{w^\top \text{swap}_1(x) \geq 0}) \, \text{sign}(\langle x, e_1 \rangle) \geq 0$, so that

$$w_1 \geq 0 \implies \langle D(w, \mathbf{0}), e_2 \rangle \leq 0.$$

Moreover if $w_1 > 0$, the above expectation is non-zero since there is at least a non-zero measure subset of $\mathbb{R}^d$ for which $(\mathbb{1}_{w^\top x \geq 0} - \mathbb{1}_{w^\top \text{swap}_1(x) \geq 0}) |x_2| > 0$. Symmetric arguments allow to derive Equation (62).

G.3 Proof of Equation (63).

Similar computations as above yield

$$\begin{aligned}\langle D(w,\mathbf{0}), w_{3:d}\rangle &= \mathbb{E}[\mathbb{1}_{w^\top x\geq 0} y(z)\langle \xi, w_{3:d}\rangle] \\ &= \frac{1}{2}\mathbb{E}[(\mathbb{1}_{w^\top(z+\xi)\geq 0} - \mathbb{1}_{w^\top(z-\xi)\geq 0}) y(z)\langle \xi, w_{3:d}\rangle] \\ &= \frac{1}{2}\mathbb{E}[y(z)\mathbb{1}_{|\xi^\top w_{3:d}|\geq |z^\top w|}|\langle \xi, w_{3:d}\rangle|] \\ &= \frac{1}{2}\mathbb{E}_\xi\mathbb{E}_z[\mathbb{1}_{y(z)=1}\mathbb{1}_{|\xi^\top w_{3:d}|\geq |z^\top w|}|\langle \xi, w_{3:d}\rangle|] \\ &\quad - \frac{1}{2}\mathbb{E}_\xi\mathbb{E}_z[\mathbb{1}_{y(z)=-1}\mathbb{1}_{|\xi^\top w_{3:d}|\geq |z^\top w|}|\langle \xi, w_{3:d}\rangle|]\end{aligned}$$

Assume w.l.o.g. that $w_1, w_2 \geq 0$. For a fixed ξ and since $y(z) = -\text{sign}(z_1 z_2)$, the region $\{\mathbb{1}_{y(z)=-1}\mathbb{1}_{|\xi^\top w_{3:d}|\geq |z^\top w|} \geq 0\}$ is smaller (w.r.t. the Gaussian measure) than the region $\{\mathbb{1}_{y(z)=1}\mathbb{1}_{|\xi^\top w_{3:d}|\geq |z^\top w|} \geq 0\}$. It is indeed a consequence of the following observation: if $y(z) = -1$ then $|z^\top w| \geq |\text{swap}_1(z)^\top w|$, i.e.,

$$\begin{aligned}&\text{if} \quad \mathbb{1}_{y(z)=-1}\mathbb{1}_{|\xi^\top w_{3:d}|\geq |z^\top w|} = 1 \\ \text{then} \quad &\mathbb{1}_{y(\text{swap}_1(z))=1}\mathbb{1}_{|\xi^\top w_{3:d}|\geq |\text{swap}_1(z)^\top w|} = 1,\end{aligned}$$

and $\text{swap}_1(z)$ also follows a standard Gaussian distribution.

As a consequence, if $w_1, w_2 \geq 0$, the previous inequality implies that $\langle D(w,\mathbf{0}), w_{3:d}\rangle \geq 0$. Moreover, if both $w_1, w_2 > 0$ and $w_{3:d} \neq 0$, the above difference becomes positive, so that $\langle D(w,\mathbf{0}), w_{3:d}\rangle > 0$. More generally, we have shown by symmetric argument that

$$\text{sign}(D(w,\mathbf{0}), w_{3:d}\rangle = \text{sign}(w_1 w_2)\mathbb{1}_{w_{3:d}\neq 0}$$

G.4 Proof of Equation (64).

Assume that $w_{3:d} = 0$, Equation (65) then yields

$$\begin{aligned}\langle D(w,\mathbf{0}), e_2\rangle &= -\frac{1}{2}\mathbb{E}_x[(\mathbb{1}_{w^\top x\geq 0} - \mathbb{1}_{w^\top \text{swap}_1(x)\geq 0})\text{sign}(x_1)|x_2|] \\ &= -\frac{\text{sign}(w_1)}{2}\mathbb{E}_x[\mathbb{1}_{|w_1 x_1|\geq |w_2 x_2|}\text{sign}(w_1 x_1)\text{sign}(x_1)|x_2|] \\ &= -\frac{\text{sign}(w_1)}{2}\mathbb{E}_x[\mathbb{1}_{|w_1 x_1|\geq |w_2 x_2|}|x_2|].\end{aligned}$$

And similarly $\langle D(w,\mathbf{0}), e_1\rangle = -\frac{\text{sign}(w_2)}{2}\mathbb{E}_x[\mathbb{1}_{|w_2 x_2|\geq |w_1 x_1|}|x_1|]$. Now note that the transformation $\tilde{x} = (x_2, x_1, x_3, \dots, x_d)$ does not change the data distribution and thus yields the following inequalities

$$\begin{aligned}\mathbb{E}_x[\mathbb{1}_{|w_1 x_1|\geq |w_2 x_2|}|x_2|] &= \mathbb{E}_x[\mathbb{1}_{|w_1 x_2|\geq |w_2 x_1|}|x_1|] \\ &\begin{cases}\leq \mathbb{E}_x[\mathbb{1}_{|w_2 x_2|\geq |w_1 x_1|}|x_1|] & \text{if } |w_2|\geq |w_1| \\ \geq \mathbb{E}_x[\mathbb{1}_{|w_2 x_2|\geq |w_1 x_1|}|x_1|] & \text{if } |w_2|\leq |w_1|.\end{cases}\end{aligned}$$

Moreover, the inequalities are strict if $|w_2| > |w_1|$ or $|w_2| < |w_1|$, so that

$$|\langle D(w, \mathbf{0}), e_2\rangle| > |\langle D(w, \mathbf{0}), e_1\rangle| \quad \text{if } |w_1| > |w_2|,$$
$$|\langle D(w, \mathbf{0}), e_2\rangle| < |\langle D(w, \mathbf{0}), e_1\rangle| \quad \text{if } |w_1| < |w_2|.$$